Europäisches Patentamt	European Patent Office	Office européen des brevets

Bescheinigung Certificate Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.	The attached documents are exact copies of the European patent application described on the following page, as originally filed.	Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr.	Patent application No.	Demande de brevet n°
	02425791.7	

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

R C van Dijk

Anmeldung Nr:
Application no.: 02425791.7
Demande no:

Anmeldetag:
Date of filing: 23.12.02
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

STMicroelectronics S.r.l.
Via C. Olivetti, 2
20041 Agrate Brianza (Milano)
ITALIE

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se referer à la description.)

Method of analysis of a table of data relating to expressions of genes and relative identification system of co-expressed and co-regulated groups of genes

In Anspruch genommene Priorität(en) / Priority(ies) claimed /Priorité(s) revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/ Classification internationale des brevets:

G06F/

Am Anmeldetag benannte Vertragstaaten/Contracting states designated at date of filing/Etats contractants désignées lors du dépôt:

AT BE BG CH CY CZ DE DK EE ES FI FR GB GR IE IT LI LU MC NL PT SE SI SK TR

Italian Text Pursuant to Art. 14.2

"METODO DI ANALISI DI UNA TABELLA DI DATI RELATIVI ALL'ESPRESSIONE DI GENI E RELATIVO SISTEMA DI IDENTIFICAZIONE DI GRUPPI DI GENI CO-ESPESSI E CO-REGOLATI"

CAMPO DELL'INVENZIONE

La presente invenzione concerne in generale l'analisi genomica e più in particolare un metodo e relativo sistema di identificazione di gruppi di geni co-regolati e co-espressi dall'analisi di dati relativi all'espressione di geni variabile nel tempo o relativa a condizioni differenti.

BACKGROUND DELL'INVENZIONE

La conoscenza della sequenza del genoma umano e di altri organismi fornisce alla ricerca strumenti d'importanza fondamentale per lo sviluppo di strategie di prevenzione e cura delle malattie, che nella maggior parte dei casi sembrerebbero dovute all'azione di più geni contemporaneamente.

Tutte le cellule di un organismo contengono lo stesso patrimonio genetico, ma il numero di geni che sono espressi, ovvero copiati in mRNA, è specifico per tipo cellulare; in questo senso lo mRNA è lo specchio dei geni attivi. Ogni cellula del nostro corpo esprime soltanto uno specifico set di geni in accordo con un programma altamente regolato che conferisce a quella cellula il suo disegno distintivo e le sue capacità funzionali. Il programma d'espressione genica di un genoma definisce il ruolo e il comportamento d'ogni cellula di un organismo.

L'informazione genetica d'ogni cellula controlla che le proteine siano sintetizzate, in quale quantità, attraverso il preciso controllo dell'espressione genica dei trascritti da ogni gene (*Regolazione*).

Il Gene-Chip rappresenta uno strumento straordinario per lo studio del profilo di espressione genica, inteso come l'insieme degli stati di attività di un set di geni.

Uno studio comparativo dell'espressione genica di diversi tessuti, o nello stesso tempo a differenti stadi di sviluppo, consente di capire quali geni distinguono un tipo cellulare da un altro e come i vari tipi cellulari si differenziano.

Grazie a queste tecnologie, oggi si ha a disposizione una consistente mole di dati relativa al livello d'espressione genica.

Un obiettivo della ricerca genomica è l'interpretazione dei profili d'espressione genica, ponendo l'attenzione su quali siano le interazioni gene–gene tra geni che concorrono ad uno stesso processo regolatorio. Dato che ogni esperimento produce una grande quantità di dati, essi devono essere organizzati utilizzando tecniche di "Data Mining".

Con il termine "Data Mining", citato in letteratura anche come "Knowledge Discovery in Database" (scoperta della "conoscenza" dai dati contenuti nei database) si indica l'insieme di quelle tecniche di estrazione di informazione potenzialmente utile e sconosciuta in precedenza, da un insieme di dati.

Note tecniche di Data Mining sono quelle basate su algoritmi di raggruppamento (clustering) classificabili in due differenti tipologie:

- Clustering semantico, basato su proprietà semantiche di una certa entità;
- Clustering numerico, basato su proprietà quantitative di una certa entità.

Mentre secondo gli approcci classici veniva studiato un gene o sequenza per volta, l'attuale tendenza genomica è verso lo sviluppo di tecnologie che permettano l'analisi parallela di molte migliaia di geni contemporaneamente. Lo stato dell'arte in questo ambito riguarda la costruzione di matrici (microarray) con acidi nucleici per lo studio dei livelli di mRNA di campioni biologici.

L'attuale stato dell'arte nell'analisi dell'informazione genomica è caratterizzato dall'uso di tecniche di raggruppamento (clustering) di tipo esclusivo e non supervisionato. Attraverso queste tecniche, le sequenze geniche, che sono rappresentate da un vettore i cui elementi sono le espressioni temporali

determinate in uno specifico contesto biologico, o più in generale i valori di espressione in diverse condizioni, sono raggruppate per similarità con algoritmi che non richiedono alcuna conoscenza a priori.

Il termine "clustering" è quindi utilizzato, in questo specifico contesto, per indicare criteri di raggruppamento con i quali si può partizionare una tabella di dati relativi a geni in sotto-tabelle contenenti dati di geni che presentano caratteristiche simili. In sostanza, con i criteri di raggruppamento ("clustering") è possibile partizionare le sequenze geniche secondo caratteristiche di similarità nell'andamento dell'espressione temporale o nella comparazione di diversi stati.

Una breve presentazione dei criteri di raggruppamento noti è effettuata nel paragrafo successivo.

DATA MINING

I dati relativi al livello di espressione di geni esaminati, vengono organizzati, mantenuti e aggiornati in una tabella, come la Tabella 1. Nelle diverse colonne sono riportati i livelli di espressione in diversi istanti temporali o per diverse condizioni iniziali del ciclo di esperimenti.

L'insieme dei livelli di espressione di un gene nei diversi istanti temporali (dati della stessa riga), è chiamato *Profilo di espressione temporale* ("pattern of gene expression").

La colonna ORF[1] contiene valori alfanumerici identificativi dei geni, detti anche "accession numbers", atti a specificare a quali geni riferirsi nei siti web esterni che vengono consultati per la ricerca. Ogni qualvolta, viene scoperto un nuovo gene, gli viene assegnato un "accession number" per poterlo introdurre nei database pubblici.

Le celle interne alla tabella contengono informazioni per uno specifico gene correlato al suo campione. Il valore "5.8" in corrispondenza della prima riga e

1 Open Reading Frame

della quarta colonna della Tabella 1, indica che il valore o livello dell'espressione genica, osservato per il gene YAL001C dopo due ore dal primo esperimento, è 5.8. La cella vuota indica assenza di informazione.

Dato un set di geni, l'insieme dei loro stati di attività, intesi come livelli d'espressione ad un dato istante (dati nella stessa colonna), è chiamato *profilo di espressione genica*.

Allo stato dell'arte, la tecnica computazionale più utilizzata per l'analisi di espressione genica è il clustering numerico di geni basato sulla similarità dei pattern di espressione.

Nell'ambito della bioinformatica, le tecniche di raggruppamento (clustering) producono gruppi di geni (CLUSTERS) per:

- Estrarre dei "motivi"[2] regolatori;
- Dedurre famiglie funzionali;
- Classificare tipi di cellule, campioni di tessuto ecc..

Estrazione dei "motivi" regolatori

Se i motivi del gene sono noti, come nel caso del lievito, è possibile identificare i motivi regolatori di un CLUSTER di geni che presentano simili livelli di espressione genica. In altre parole è possibile dedurre la co-regolazione dalla co-espressione: è probabile cioè che geni co-espressi, facenti parte dello stesso CLUSTER, contengano motivi regolatori comuni.

I CLUSTER, ottenuti raggruppando sequenze temporali di espressioni geniche simili, possono essere analizzati per predire i fattori di trascrizione responsabili della sintesi di una proteina.

Deduzione di famiglie funzionali

Un altro obiettivo delle tecniche di raggruppamento (clustering) è quello di

identificare geni che abbiano simili funzioni ovvero che intervengono in uno stesso processo cellulare. Se ad esempio un gene sconosciuto è raggruppato con un numero di geni la cui famiglia funzionale è nota (per mezzo del clustering semantico), è possibile dedurre la funzione del gene in esame.

Classificazione di tipi di cellule, campioni di tessuto ecc.

Un altro modo di utilizzare dati di espressione genica è la classificazione intesa come identificazione e distinzione di diversi tipi di cellule, campioni di tessuto ecc..

Si ipotizza che geni appartenenti ad uno stesso gruppo (CLUSTER) concorrono alla formazione di uno stesso tipo di cellula o di uno stesso tessuto.

Le informazioni ottenute con tecniche di Data Mining, possono essere interpretate con l'ausilio di rappresentazioni grafiche come quelle nelle Figure 1 e 2. Gli andamenti temporali dei geni nel grafico di Figura 1 sono riferiti ad uno stesso campione (profili di espressione temporale).

Come si può osservare, i geni 1 e 2 presentano delle caratteristiche simili, poichè hanno entrambi un analogo andamento temporale (valori del pattern di espressione molto "vicini"). Ci si aspetta dunque, che questi due geni rispondano entrambi a stessi stimoli interni o esterni: per esempio, la loro attività potrebbe essere influenzata dallo stesso set di ingressi, ovvero dallo stesso set di geni regolatori.

Diversamente, nel grafico di Figura 2 sono confrontati livelli di espressione genica ad un istante temporale fissato, ma relativi a campioni differenti. Questo confronto è d'aiuto per capire se un gene (magari in concomitanza con altri geni) è responsabile di una certa malattia.

Se ad esempio un campione (T5) di un portatore sano di una malattia, quali ad esempio il diabete, ha un gene (Gene1) con un livello di espressione che presenta

---

2 Brevi sequenze di DNA che si legano all'RNA polimerase definendo il punto di inizio della trascrizione.

una forte variazione rispetto al valore dello stesso gene esaminato da altri campioni di soggetti non affetti da diabete, si potrebbe ipotizzare che il gene Gene1 sia la causa della malattia in questione.

Le tecniche di raggruppamento (clustering), oltre ad essere un potente mezzo per l'analisi dell'espressione genica, sono utilizzate nella cosiddetta fase di pre-processing di una rete dinamica regolatoria di geni (Gene Network). Per modellizzare una rete dinamica regolatoria è necessario, come primo passo, conoscere da quali e quanti geni essa è costituita. Nella fase di pre-processing vengono selezionati uno o più gruppi di geni che possono costituire una Gene Network.

Data l'importanza delle tecniche di raggruppamento (clustering), nei paragrafi successivi sono presentate queste tecniche di Data Mining.

Clustering di tipo numerico

Una tecnica di clustering di tipo numerico fa uso d'algoritmi matematici per raggruppare geni basandosi sulla similarità dei valori d'espressione genica ricavati.

Brevemente, una tecnica di clustering numerico si compone di due parti: una *misura di distanza* che indica quanto siano simili i pattern d'espressione di una coppia di geni (o più genericamente di due CLUSTER) e un algoritmo di clustering per identificare CLUSTER di simili pattern d'espressione genica basati sulla misura di distanza scelta.

Clustering di tipo semantico

Il sequenziamento del genoma umano e di vari altri organismi modello in questi ultimi anni ha dato una particolare risonanza a quel settore della bioinformatica che si occupa dello studio del DNA e delle proteine. Grazie, quindi, all'introduzione dei metodi di sequenziamento automatizzato del DNA e ai molti progetti di sequenziamento del genoma, la quantità d'informazione su sequenze proteiche è aumentata a dismisura. La gran quantità di dati così prodotti necessita

di essere collezionata, memorizzata e distribuita.

Poichè la gestione di tale mole di dati richiede un uso intensivo del computer, lo sviluppo dei database è uno dei punti di forza del progetto genoma. I database devono essere progettati accuratamente e la loro architettura deve essere tale da contenere informazioni sulla mappa (per esempio la locazione fisica di un gene), sulle sequenze (nucleotidiche e proteiche) e devono anche fornire dei collegamenti a database contenenti informazioni di carattere scientifico e medico.

Oggi tramite Internet sono facilmente accessibili una serie di banche dati in cui i laboratori di genetica di tutto il mondo riversano quotidianamente i dati da loro prodotti, sviluppando inoltre strumenti per l'analisi e il confronto di queste informazioni. I maggiori database pubblici che memorizzano sequenze nucleotidiche sono: GenBank (http://www.ncbi.nlm.nih.gov/), EMBL (European Molecular Biology Laboratory, http://www.ebi.ac.uk/Information/index.html) e DDBJ (DNA DataBase of Japan, http://www.ddbj.nig.ac.jp). I più comuni database che memorizzano sequenze proteiche sono: PIR (Protein Identification Resource – National Biomedical Research Foundation), Swissprot e GenPept (entrambi distribuiti con GenBank). In aggiunta alle informazioni sulle sequenze, essi contengono informazioni sui motivi regolatori delle proteine e sulle altre caratteristiche della struttura delle proteine.

In generale, questi database contengono informazioni circa le sequenze geniche note, quali ad esempio informazioni riguardanti i "domini funzionali" o attributi (le cosiddette "ontologies") di uno specifico prodotto del gene (gene product). Esempi di attributi di un prodotto del gene sono la *funzione molecolare*[3] (molecular function), il *processo biologico*[4] (biological process) e la *componente*

[3] Con il termine *funzione molecolare* si indica la capacità che un gene potenzialmente ha, l'attività biochimica del prodotto del gene, ciò che esso può fare senza specificare dove o quando questo avviene. Esempi di tali termini sono "enzyme", "transporter", "ligand", 2adenylate cyclase" o "Toll receptor ligand".

[4] Il termine *processo biologico* si riferisce all'obbiettivo biologico a cui il gene o il prodotto del gene contribuisce. Esempi di tali termini sono "cell growth and maintenance", "signal transduction", "pyrimidine metabolism" o "cAMP biosynthesis".

cellulare[5] (cellular component).

Questi attributi di un gene sono parametri di natura semantica, cioè espressi da parole scelte in un vocabolario controllato. La creazione di un simile vocabolario è un obiettivo del Gene Ontology Consortium (http://www.geneontology.org, http://genome-www.stanford.edu/saccharomyces/help/GO.html).

Associando a questi parametri semantici corrispondenti valori numerici è possibile applicare note tecniche di clustering per raggruppare geni che condividono simili caratteristiche semantiche. In questo modo si ottengono CLUSTERS di geni che presentano domini funzionali simili o attributi (ontologies) simili.

In Figura 3 viene mostrato un esempio di report ottenuto eseguendo una query al sito LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/). La query è inoltrata inserendo l'accession number (identificativo univoco per la base dati specificata) del gene d'interesse (nell'esempio mostrato l'accession number è D50497). Tra i risultati della query ci sono anche le ontologies, indicate nel riquadro ove è presente il marchio Gene Ontology™ (marchio del Gene Ontology Consortium).

Per meglio comprendere il significato dell'espressione "dominio funzionale", si fa una breve digressione sulla struttura delle proteine che costituiscono il DNA.

LA STRUTTURA DELLE PROTEINE

Nelle proteine allo stato nativo, le catene polipeptidiche non hanno una struttura tridimensionale disordinata o casuale, bensì, per ciascuna proteina, sono tutte disposte nello spazio nello stesso modo e si presentano come oggetti con una forma (struttura) identica in tutte le molecole di una data proteina. Ciò dipende dal fatto che ogni molecola, inserita in un determinato ambiente, assume una disposizione nello spazio che consenta di stabilire il massimo numero possibile di interazioni sia tra atomi o gruppi di atomi che fanno parte della stessa molecola,

[5] Il termine *componente cellulare* si riferisce si riferisce al sito nella cellula in cui un prodotto del gene è attivo.

sia tra atomi o gruppi di atomi delle molecole vicine.

Si è adottata la convenzione di distinguere diversi livelli di complessità strutturale, descrivendo una struttura primaria, una struttura secondaria, una struttura terziaria e, in alcuni casi, una struttura quaternaria delle proteine. Va sottolineato, però, che si tratta di un artificio descrittivo, in quanto i diversi livelli di complessità strutturale si integrano per dare origine a qualcosa di unitario, che è appunto la struttura tridimensionale specifica di ciascuna proteina nel suo complesso.

La successione delle unità di amminoacidi prende il nome di struttura primaria e con essa si indica esattamente lo scheletro covalente della catena peptidica. Ogni proteina è caratterizzata da una propria specifica sequenza di aminoacidi, diversa da quella di ogni altra proteina. Il DNA, quindi, contiene in codice le informazioni riguardanti la struttura primaria di tutte le proteine.

È di fondamentale importanza, sapere anche come le catene peptidiche siano disposte ed associate nello spazio per capire come una proteina svolga determinate funzioni.

La struttura terziaria, che rappresenta la struttura tridimensionale vera e propria della proteina, è diversa da proteina a proteina e consiste nel modo con cui la catena, in parte organizzata in struttura secondaria, si raggomitola per dare origine alla proteina nativa. Le porzioni di catene in strutture secondarie sono collegate da tratti ad andamento irregolare che formano anse (loops), in alcuni casi dotate di una certa mobilità. Il raggomitolamento non è casuale ed è fortemente condizionato dall'ambiente in cui la proteina si trova. Il raggomitolamento della catena polipeptidica ha l'importante conseguenza di portare vicine le une alle altre le catene laterali di aminoacidi situati in punti lontani della struttura primaria.

Questi "sottogomitoli" prendono il nome di "domini strutturali" della proteina. Spesso essi mantengono la loro struttura anche quando vengono separati dal resto della proteina.

A volte i *domini strutturali* svolgono una specifica sotto-funzione nell'ambito

della funzione biologica della proteina (ad esempio, in certi enzimi un dominio svolge la funzione catalitica, mentre un altro è deputato a interagire con le sostanze che regolano l'attività dell'enzima stesso): in questi casi, una relativa autonomia strutturale di una parte della catena polipeptidica è associata ad una relativa autonomia funzionale. Questi domini strutturali vengono indicati come *domini funzionali* della proteina. È importante osservare che in un certo numero di proteine, che svolgono funzioni parzialmente simili, sono presenti domini simili che svolgono in tutte la stessa funzione.

Gli algoritmi più diffusi ed utilizzati per il raggruppamento sono l'algoritmo di clustering gerarchico agglomerativo, il K-means e il SOM (Self Organizing Map). I risultati degli algoritmi di clustering utilizzati dipende dalla metrica utilizzata per definire il criterio di similarità tra sequenze geniche. Di conseguenza, due sequenze geniche che sono giudicate simili usando una certa metrica, possono essere giudicate molto diverse tra loro utilizzando una metrica differente.

Attualmente la ricerca genomica è limitata all'uso degli algoritmi di clustering per l'individuazione di sequenze geniche che si comportano in maniera similare nel contesto di un determinato processo biologico.

Un limite delle tecniche note di individuazione di gene networks è rappresentato dal fatto che non è possibile individuare sequenze geniche applicando più criteri contemporaneamente pesati opportunamente al fine di determinare sequenze che possono essere anche differenti tra loro dal punto di vista dell'espressione genica nel tempo, ma che godono di proprietà specifiche di particolare interesse.

SCOPO E SOMMARIO DELL'INVENZIONE

È stato trovato ed è l'oggetto della presente invenzione un metodo per l'analisi automatica dell'informazione genomica, al fine di identificare relazioni tra geni che concorrono ad uno stesso processo regolatorio. Il metodo dell'invenzione consente di determinare relazioni complesse tra geni, che vanno oltre le semplici operazioni di clustering dei metodi noti che mirano alla determinazione di geni co-espressi o co-regolati.

Il metodo dell'invenzione si applica ad una tabella di dati relativi all'andamento dell'espressione genica nel tempo o relativa a condizioni di stress differenti, e non dipende dal metodo utilizzato per ricavare tale tabella.

Dapprima si sceglie un criterio di raggruppamento (clustering) e lo si applica alla tabella ricavando delle sotto-tabelle di dati relativi a gruppi di geni (CLUSTERS) che soddisfano il criterio di clustering scelto.

Si generano quindi tutte le possibili combinazioni di coppie di sotto-tabelle e si calcolano dei parametri caratteristici per i geni contenuti in tali sotto-tabelle. Infine si calcola per ogni combinazione un valore caratteristico con un algoritmo di decisione definito in funzione di tali parametri, considerando i geni della combinazione come costituenti una 'Gene Network' se questo valore caratteristico eccede una soglia predefinita.

Preferibilmente, si sceglie anche un insieme di criteri logici di filtraggio dei dati della tabella, generando altre sotto-tabelle di dati di gruppi di geni che soddisfano il rispettivo criterio logico e si calcolano le combinazioni tra coppie di sotto-tabelle, ottenute con i criteri logici o di raggruppamento.

Preferibilmente, l'algoritmo di decisione è un algoritmo fuzzy i cui antecedenti e conseguenti sono definiti in funzione di questi parametri caratteristici.

Il metodo dell'invenzione è implementato da un relativo sistema di identificazione di gruppi di geni co-espressi e co-regolati. Il cuore di questo sistema di identificazione è un sotto-sistema intelligente che elabora i parametri caratteristici di gruppi di geni e produce in uscita dati di gruppi di geni identificati come 'Gene Networks'.

Preferibilmente, tale sotto-sistema intelligente è un sotto-sistema fuzzy addestrato off-line identificato mediante una rete neuronale.

L'invenzione è più precisamente definita nelle annesse rivendicazioni.

BREVE DESCRIZIONE DEI DISEGNI

I diversi aspetti e vantaggi dell'invenzione risulteranno ancor più evidenti attraverso una descrizione dettagliata facendo riferimento ai disegni allegati, in cui:

la **Tabella 1** è un esempio di tabella di dati relativi all'espressione di geni;

la **Figura 1** è un diagramma dei livelli di espressione di geni in diversi istanti temporali relativamente ad uno stesso campione di DNA;

la **Figura 2** è un diagramma interpolato dei livelli di espressione di geni ad un istante prefissato relativamente a differenti campioni (T1, ..., T6) di DNA;

la **Figura 3** è un esempio di rapporto ottenuto eseguendo una query al sito LocusLink;

la **Figura 4** mostra una forma di realizzazione preferita di un sistema dell'invenzione;

la **Figura 5** mostra dei possibili diagrammi di dispersione;

la **Figura 6** mostra degli esempi di diagrammi di dati correlati secondo una legge quadratica;

la **Figura 7** mostra possibili andamenti temporali di sequenze geniche;

la **Figura 8** mostra un set di dati per l'addestramento del sistema Fuzzy dell'invenzione;

la **Tabella 2** mostra un'insieme di valori di espressione di geni del lievito *S. cerevisiae* in diversi istanti;

la **Tabella 3** mostra delle informazioni ricavabili da Saccharomyces Genome Database per dei geni riportati in Tabella 2;

la **Tabella 4** mostra dati relativi a dei geni della Tabella 2 che sono stati raggruppati in un CLUSTER;

la **Tabella 5** mostra possibili combinazioni tra gruppi di geni e il valore caratteristico associato a ciascuna combinazione;

le **Tabelle 6 e 7** mostrano dati relativi a geni raggruppati nei CLUSTERS 26 e 30;

la **Tabella 8** riporta livelli di espressione di geni della combinazione tra i CLUSTERS delle Tabelle 6 e 7;

la **Tabella 9** mostra i livelli di espressione della Tabella 8 normalizzati tra 0 e 1;

la **Tabella 10** mostra valori degli incrementi dei livelli di espressione della

Tabella 9.

DESCRIZIONE DI ALCUNE FORME DI REALIZZAZIONE DELL'INVENZIONE

Il metodo della presente invenzione consente di individuare gruppi di geni ('Gene Network') probabilmente coinvolti in un processo regolatore. Tale metodo è basato su un algoritmo di decisione che, diversamente dai metodi noti, identifica gruppi di geni co-espressi o co-regolati utilizzando contemporaneamente sia criteri di clustering che criteri logici di filtraggio. Da ciascun gruppo di geni così ottenuto si ricavano dei parametri caratteristici e, con un algoritmo di decisione basato su tali parametri caratteristici, si calcola un valore caratteristico: se tale valore caratteristico eccede una certa soglia prestabilita, allora il relativo gruppo di geni è identificato come 'Gene Network', altrimenti viene scartato.

Il notevole vantaggio di questa tecnica consiste nel fatto che vengono superati i limiti dei metodi attuali basati esclusivamente sul raggruppamento (clustering), consentendo di identificare un gruppo di geni come una 'Gene Network' sulla base di più criteri variamente combinati.

Preferibilmente, tale algoritmo di decisione è un algoritmo fuzzy configurato in modo tale da individuare correlazioni tra geni all'interno di una grossa mole di dati, corrispondenti all'espressione genica variabile nel tempo o relativa a condizioni differenti di sequenze di geni immobilizzate su microarray.

La rappresentazione schematica di una forma di realizzazione preferita di un sistema implementante il metodo dell'invenzione è riportata nella Figura 4.

Sono presenti tre sotto-sistemi:

1. Pre-elaborazione (CLUSTERING, FILTERING), che genera gruppi di tabelle usando criteri di raggruppamento e criteri logici di filtraggio.

2. Elaborazione (GENERAZIONE COMBINAZIONI, ESTRAZIONE CARATTERISTICHE), che genera gruppi di geni candidati Gene Networks combinando coppie di sotto-tabelle ed estraendo parametri caratteristici per ogni combinazione di geni.

3. Sotto-sistema intelligente (SISTEMA NEURONALE, SISTEMA FUZZY, THRESHOLD), addestrato off-line, che produce in uscita gruppi di geni identificati come Gene Networks.

Il sotto-sistema intelligente è preferibilmente basato sulla logica Fuzzy, opportunamente addestrato off-line, in grado di attribuire ad ogni gruppo di geni candidati un valore caratteristico mediante un algoritmo decisionale basato su Soft Computing: se questo valore caratteristico eccede una soglia prestabilita THRESHOLD, allora il relativo gruppo di geni è identificato come costituente una Gene Network.

Clustering e Filtering

Il sotto-sistema di pre-elaborazione (pre-processing) genera gruppi similari di sequenze geniche usando criteri di raggruppamento (clustering) e criteri logici di filtraggio. Ci sono diversi criteri di raggruppamento noti in letteratura, come quelli qui di seguito elencati:

- Gerarchico Agglomerativo;
- Non gerarchico Kmeans;
- Gerarchico Kmeans sequenziale;
- Non gerarchico SOM;
- Non esclusivo Fuzzy Clustering.

Per ogni gene si riportano in ingresso m valori di espressione genica, relativi ad m esperimenti condotti in istanti temporali differenti o condizioni differenti. Il sistema genera un certo numero di gruppi di geni (CLUSTER) secondo il criterio utilizzato e le impostazioni scelte per l'esecuzione.

Ai fini dello studio delle Gene Networks, è interessante selezionare gruppi di geni che presentano altre caratteristiche, oltre la similarità dei profili di espressione temporale. Questo è reso possibile mediante tecniche di filtraggio che selezionano gruppi di geni, in base al valore assunto da uno o più attributi del gene stesso.

La scelta dei criteri da utilizzare deve essere eseguita considerando l'idoneità dei criteri di clustering nel costituire gruppi di sequenze geniche simili. Ad esempio, se si vuole verificare l'influenza di gruppi estesi tra loro e verso singoli geni, è consigliabile usare un criterio logico di filtraggio stringente e algoritmi di clustering che generano gruppi estesi di CLUSTERS. Una simile scelta può consistere in un metodo gerarchico accoppiato alla metrica di aggiornamento della matrice della distanze di tipo single linkage.

I criteri di filtraggio sono tutti quei criteri logici che un utente può impostare sui dati di una tabella. Ad esempio, un criterio di filtraggio può consistere nel selezionare tutti quei geni il cui livello di espressione eccede un certo valore all'inizio dell'esperimento. Un altro criterio di filtraggio può essere quello di considerare quei geni il cui livello di espressione dopo 7 minuti dall'inizio dell'esperimento sia compreso in un certo intervallo e dopo 14 minuti sia compreso in un altro intervallo, ecc.

Generazione di combinazioni e composizione di gruppi di geni candidati Gene Network

Si supponga di aver generato K sotto-tabelle di geni (CLUSTER) con un criterio di raggruppamento e M sotto-tabelle di geni (FILTER) con criteri logici di filtraggio. Secondo il metodo dell'invenzione, si generano tutti i possibili gruppi di geni ottenuti combinando le sotto-tabelle a coppie:

1. $\binom{K}{2} = \frac{K!}{(K-2)!*2!} = \frac{K*(K-1)}{2}$ combinazioni di tipo CLUSTER-CLUSTER;
2. $\binom{M}{2} = \frac{M!}{(M-2)!*2!} = \frac{M*(M-1)}{2}$ combinazioni di tipo FILTER-FILTER;
3. $K*M$ combinazioni di tipo CLUSTER-FILTER.

Di queste combinazioni, preferibilmente sono essere scartate quelle che danno origine a gruppi con un numero di geni minore di una soglia prestabilita e quelle che danno origine a gruppi già formati da una precedente combinazione.

Ogni gene della combinazione può essere referenziato con una stringa che ne indica il gruppo di provenienza. Ad esempio, un gene è etichettato con C2 se il gruppo di provenienza è il cluster 2. In una combinazione del tipo FILTER-FILTER o CLUSTER-FILTER, un gene presente in entrambi i gruppi è etichettato con F_iF_j o C_iF_j, dove i pedici *i* e *j* indicano gli indici del CLUSTER o del FILTER di provenienza.

Qualora si voglia esaminare come il comportamento di un certo gene influenzi un intero CLUSTER, è preferibile generare combinazioni tra un FILTER costituito solo da quel gene e gruppi (CLUSTERS) costituiti da più geni.

Estrazione delle caratteristiche

La fase più significativa è costituita dall'estrazione delle caratteristiche dal momento che esse stesse indicano la tipologia di correlazioni che si vuole individuare. Secondo un aspetto innovativo della presente invenzione, si usano parametri di natura numerica, legati al profilo di espressione genica, e parametri che invece rappresentano un contenuto semantico, o parametri misti, ottenuti come combinazione di entrambi gli elementi. Un esempio di parametro di natura semantica è la rappresentazione numerica dei domini funzionali della sequenza aminoacidica omologa corrispondente alla sequenza nucleotidica di partenza immobilizzata su microarray.

Un parametro di natura semantica considerato nell'analisi del gene network è, ad esempio, la percentuale di geni della combinazione con lo stesso dominio funzionale. A questa percentuale corrisponde un numero tra 0 e 1. Se il valore del parametro è unitario, tutti i geni della combinazione in esame hanno lo stesso dominio funzionale. Se il valore è nullo, i geni non hanno nessun dominio in comune, mentre in tutti gli altri casi il parametro assume un valore compreso tra 0 e 1.

Un altro parametro di natura semantica, analogamente al caso precedente, è relativo alla percentuale di geni che presentano attributi (ontologies) uguali o appartenenti ad una stessa categoria. È intuitivo il fatto che potrebbero essere

considerati altri parametri semantici estendendo questa analisi ad altre caratteristiche semantiche delle sequenze geniche. Va inoltre sottolineato che questi parametri si riferiscono a caratteristiche di natura semantica ma sono espressi in forma numerica.

Secondo una forma preferita di realizzazione del metodo dell'invenzione, sono usati sei parametri di natura numerica P1, ..., P6. Ogni parametro ha un range di variazione tra zero e uno.

Il primo parametro P1 è uguale al modulo del coefficiente di correlazione lineare tra le espressioni di coppie di geni della stessa combinazione se la correlazione è positiva, altrimenti è nullo. Il secondo parametro P2 è analogo a P1, ma è nullo se la correlazione lineare è positiva. Il terzo parametro P3 indica il valore della correlazione quadratica della combinazione. Quanto più prossimo a uno è il valore della correlazione, tanto più i geni della combinazione sono correlati.

Il quarto parametro P4 indica la percentuale di geni del gruppo che ha il valore di espressione genica finale (cioè l'ultimo attributo del gene) maggiore o minore del valore di espressione genica iniziale (primo attributo). In pratica, si calcola la percentuale di geni che ha lo stesso comportamento dal punto di vista della variazione complessiva.

Il quinto parametro P5 indica la percentuale di geni del gruppo che ha lo stesso andamento temporale (crescente o decrescente). Infine, l'ultimo parametro P6 indica la percentuale di geni che presenta un'escursione massima (picco) nello stesso istante temporale.

I parametri introdotti hanno lo scopo di verificare se il gruppo di geni in esame è costituito da geni espressi differentemente che partecipano ad uno stesso processo regolatorio e pertanto se le relazioni tra gli stessi possano essere modelizzate mediante una rete dinamica regolatoria (Gene Network).

Il fatto di usare questi sei parametri al fine di determinare gruppi di geni co-espressi e co-regolati permette di avere un metodo di identificazione robusto e

capace di un discernimento multiobiettivo. È necessario ricordare che l'approccio può essere generalizzato a qualsiasi parametro d'interesse che esprima una correlazione di qualsiasi natura tra valori di espressione e geni.

Inoltre è possibile usare parametri che possono avere o completamente un significato biologico di natura semantica oppure a parametri misti e più complessi che esprimono contemporaneamente una relazione di correlazione numerica e una di correlazione semantica. In quest'ultimo caso si deve far uso di database esterni che possono essere interrogati e di un'elaborazione dei dati ritornati per una loro codifica numerica che esprime l'eventuale correlazione semantica.

Nel seguito, si esamina in dettaglio il significato dei sei parametri proposti.

Parametri relativi alla correlazione (P1, P2, P3)

La correlazione indica il grado di relazione tra geni. Per mezzo di essa si cerca di determinare quanto bene un'equazione lineare o un'altra equazione qualsiasi descrivono o spiegano tale relazione.

Se con X e Y si indicano le due profili di espressione temporali o geniche da esaminare, si può costruire in un sistema di coordinate cartesiane un *diagramma a dispersione*. Se tutti i punti del diagramma a dispersione giacciono intorno ad una retta, la correlazione è detta lineare. In tal caso l'equazione che lega le due variabili è un'equazione lineare:

$$Y=a+bX \tag{1}$$

Se Y tende a crescere al crescere di X, la correlazione è detta positiva o diretta. Se Y tende a decrescere al crescere di X, la correlazione è detta negativa o inversa. Se non c'è alcuna relazione lineare tra le due sequenze, si dice esse che sono *incorrelate*. Il grado di correlazione lineare tra due sequenze geniche è dato dal coefficiente di correlazione lineare così definito:

$$\rho = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum \left(Y - \overline{Y}\right)^2}} \qquad -1 \le \rho \le 1 \qquad (2)$$

in cui la sommatoria va da 1 a m (con m numero di livelli di espressione calcolati per ciascun gene) e $\overline{X} = \frac{\sum X}{m}$ e $\overline{Y} = \frac{\sum Y}{m}$ rappresentano i valori medi.

La correlazione lineare, è massima quando il coefficiente ρ in modulo è uguale a 1 (il segno dipende dal fatto che una variabile cresca o diminuisca al crescere dell'altra). La Figura 5 illustra degli esempi di diagrammi a dispersione.

Un valore nullo del coefficiente di correlazione lineare implica soltanto l'assenza di una correlazione di tipo lineare, tuttavia due sequenze possono essere fortemente dipendenti e non presentare una forte correlazione di tipo lineare. Un caso tipico è quello dei punti nel piano distribuiti lungo una circonferenza.

La correlazione tra due sequenze geniche può talvolta essere di tipo quadratico, cioè la relazione che lega X e Y è l'equazione di una parabola:

$$Y=a+bX+cX^2 \qquad (3)$$

in cui a è una costante, b è il coefficiente di accrescimento lineare e c rende conto della curvatura rapportando Y ai quadrati di X.

La Figura 6 mostra degli esempi di correlazione quadratica.

In generale, qualunque sia la relazione che lega X e Y, si definisce coefficiente di correlazione, la quantità:

$$r = \pm \sqrt{\frac{\sum \left(Ystim - \overline{Y}\right)^2}{\sum \left(Y - \overline{Y}\right)^2}} \qquad (4)$$

Dove *Ystim* è il valore interpolato di Y mediante il metodo dei minimi quadrati. Si osservi che r è una quantità non dimensionale, cioè non dipende dall'unità di misura impiegata. Se la relazione tra X e Y è lineare, r coincide con il coefficiente

di correlazione lineare, altrimenti assume un significato più generale. Inoltre, nel caso in cui la relazione sia lineare,

$$r_{XY}=r_{YX}$$

cioè la quantità r è la stessa indipendentemente dal fatto che X oppure Y sia la variabile indipendente. In generale

$$r_{XY}\neq r_{YX}$$

Per quanto riguarda i parametri estratti, come si è già detto, i primi tre si riferiscono appunto alla correlazione lineare e a quella quadratica. Si supponga ora di considerare una delle combinazioni generate e si assuma che il corrispondente gruppo sia costituito da n geni. Quando si ha a che fare con un numero di sequenze geniche maggiore di due, in luogo del coefficiente di correlazione lineare si considera una matrice di correlazione lineare R così definita:

$$R=\begin{vmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{vmatrix} \qquad (5)$$

essendo ρ_{ij} il coefficiente di correlazione tra le sequenze del gene i con il gene j. Ovviamente il coefficiente di correlazione di una sequenza genica con sé stessa è uguale uno, ovvero $\rho_{ii}=1$ per ogni $i=1,\ldots, n$.

Se per ogni $i\neq j$ risulta $\rho_{ij}=0$, le n sequenze geniche sono incorrelate. In questo caso il determinante della matrice R vale 1, mentre in generale esso è compreso tra 0 e 1. Tenendo conto che $\rho_{ii}=1$ e che $\rho_{ij}=\rho_{ji}$, il numero di coefficienti calcolati è pari a:

$$\binom{n}{2}=\frac{n!}{(n-2)!*2!}=\frac{n*(n-1)}{2} \qquad (6)$$

Calcolati questi valori, si suddivide l'intervallo di valori tra 0 e 1 in sotto-intervalli, ad esempio in cinque sotto-intervalli uguali di ampiezza 0.2 e si conta il numero dei coefficienti che cade in ciascun sotto-intervallo. Ad ogni sotto-intervallo, inoltre, è associato un valore di correlazione, ad esempio pari, rispettivamente, a 0.1,0.3,0.5,0.7,0.9. Se in uno dei cinque sotto-intervalli cade un numero di coefficienti maggiore del 50% del totale, al primo parametro è assegnato il valore di correlazione corrispondente a quell'intervallo. Nel caso in cui, invece, i coefficienti siano distribuiti in maggioranza fra due intervalli, al primo parametro viene assegnato un valore che è pari alla media fra tali due intervalli.

Si assuma ad esempio che n_1 coefficienti cadano nel sotto-intervallo al quale è associato un valore di correlazione v_1, ed n_2 coefficienti cadano nel sotto-intervallo con valore di correlazione v_2. Si assuma, inoltre, che $(n_1+n_2)>50\%$ del numero totale dei coefficienti distribuiti tra i cinque intervalli. Il valore assegnato al primo parametro P1 è dato da:

$$P1 = \frac{(v_1 * n_1 + v_2 * n_2)}{(n_1 + n_2)} \tag{7}$$

Infine, nel caso in cui, la maggioranza dei coefficienti è distribuita in più di due intervalli, al primo parametro viene assegnato il valore medio di tutti i coefficienti. Nel calcolo del primo parametro si considerano solo i coefficienti $\rho_{ij}>0$.

Per il calcolo del secondo parametro P2, relativo alla correlazione lineare negativa, si procede in maniera analoga al caso precedente considerando però i coefficienti $\rho_{ij}<0$ e suddividendo in cinque intervalli uguali, l'intervallo compreso tra -1 e 0.

Per quanto riguarda l'estrazione del terzo parametro P3, per il calcolo dei coefficienti di correlazione si è considerata la forma più generale del coefficiente di correlazione data dalla (4). Tenendo conto che $r_{XY}\neq r_{YX}$, il numero dei

coefficienti da calcolare, nel caso di una combinazione con n sequenze geniche, è:

$$n^*(n\text{-}1) \tag{8}$$

Il calcolo di r implica la conoscenza di *Ystim*, ovvero del valore interpolato di Y mediante il metodo dei minimi quadrati. La parabola dei minimi quadrati interpolante l'insieme dei punti (X_i, Y_i), con i=1, ..., n è espressa dall'equazione (3):

$$Y=a+bX+cX^2 \tag{3}$$

dove i coefficienti a, b e c vengono determinati risolvendo simultaneamente le tre equazioni:

$$\begin{cases} \sum Y = aN + b\sum X + c\sum X^2 \\ \sum XY = a\sum X + b\sum X^2 + c\sum X^3 \\ \sum X^2Y = a\sum X^2 + b\sum X^3 + c\sum X^4 \end{cases} \tag{9}$$

dette *equazioni normali della parabola dei minimi quadrati*.

Noti i valori delle costanti, e sostituendoli nella (3), si ricava il valore *Ystim* e quindi il valore di r.

Al terzo parametro viene assegnato il valore medio degli $n^*(n\text{-}1)$ coefficienti di correlazione così calcolati.

Ci si aspetta che le combinazioni del tipo CLUSTER-CLUSTER abbiano valori di correlazione piuttosto elevati, in quanto, già di per sé, il clustering seleziona gruppi di geni ad alta correlazione. Ciò non toglie, che anche combinazioni del tipo CLUSTER-FILTER e FILTER-FILTER possano avere valori elevati di correlazione. In generale però il parametro di correlazione dà indicazioni più complete rispetto a quelle fornite dal clustering. Per chiarire questo aspetto, si consideri il seguente esempio: si supponga di considerare due sequenze geniche X e Y costituite da tre valori di espressione temporale, X=[1; 5; 7] e Y=[10; 50; 70]. La relazione che lega X e Y è Y=10X e pertanto il coefficiente di correlazione lineare è unitario.

Tuttavia i criteri di clustering non mettono in evidenza questo tipo di relazione. La maggior parte delle tecniche di clustering implementate, infatti, fa uso di metriche di distanza. Due sequenze geniche con valori di espressione genica molto simili sono raggruppate nello stesso CLUSTER in quanto identificano punti nello spazio *m*-dimensionale molto vicini.

Nell'esempio citato, invece, pur esistendo una relazione lineare tra le due sequenze, queste identificano punti dello spazio distanti tra loro e quindi, probabilmente, non appartenenti ad uno stesso CLUSTER. L'unico criterio di clustering che fa eccezione a questa regola, è il metodo agglomerativo che fa uso del coefficiente di Pearson. Infatti, questa metrica è una misura di similarità e non di distanza e non soddisfa le proprietà metriche.

Parametri relativi al profilo di espressione

Gli ultimi tre parametri estratti, P4, P5 e P6 riguardano la similitudine tra le sequenze geniche in termini di profilo di espressione variabile nel tempo o nelle condizioni. In particolare, si considera il segno della variazione complessiva, il tipo di andamento (crescente o decrescente) e la presenza di escursioni massime in corrispondenza dello stesso istante temporale.

Il quarto parametro indica la percentuale di geni che si comportano in maniera simile dal punto di vista della variazione complessiva del valore di espressione genica.

Per ogni sequenza genica della combinazione in esame, si calcola la variazione tra il valore di espressione genica finale (cioè relativa all'ultimo attributo) e quella iniziale (relativa al primo attributo), preferibilmente tenendo conto solo del valore assoluto della variazione prescindendo dal segno. Noto il numero di sequenze geniche che ha un valore di espressione genica finale maggiore di quello iniziale, si calcola la percentuale di sequenze che ha una variazione positiva. A questa percentuale corrisponde un numero compreso tra zero e uno che viene assegnato come quarto parametro.

In pratica, più la percentuale di geni, aventi la stessa variazione, si avvicina al 50%, più il valore del quarto parametro P4 si avvicina a 0. Più invece la percentuale si avvicina al 100%, più il parametro P4 si avvicina a 1, in quanto la maggior parte dei geni del gruppo si comporta allo stesso modo.

Nel caso in cui la percentuale è bassa e prossima a 0,il valore del parametro è elevato e tende a 1. Ciò è dovuto al fatto che percentuali basse di geni aventi una variazione positiva, implicano percentuali elevate di sequenze con variazione negativa. Questo parametro mira, infatti, all'individuazione di gruppi di geni con un comportamento simile dal punto di vista della variazione complessiva del valore di espressione genica, prescindendo dal segno della variazione. Infine, percentuali del 70%, o equivalentemente del 30%, danno origine a valori del parametro prossimi a 0.5.

Tuttavia, va considerato il fatto che una sequenza genica con un valore di espressione genica finale maggiore di quello iniziale non necessariamente presenta un andamento temporale crescente, e viceversa, una variazione negativa non implica un andamento decrescente. Per identificare una gene network, è importante identificare, geni che presentino un simile andamento temporale, crescente o decrescente, prescindendo dai valori dei singoli attributi. A tal proposito si consideri la Figura 7.

Le tre sequenze geniche esemplificate A, B, C, hanno un andamento crescente, nonostante il fatto che gli andamenti sono completamente diversi e i valori di espressione genica differenti tra loro. Inoltre la sequenza A, pur avendo un andamento complessivamente crescente, ha una variazione negativa tra il valore finale e quella iniziale. Questo dettaglio sfugge ai criteri di clustering e non viene evidenziato dai quattro parametri introdotti in precedenza.

Per tale motivo è stato introdotto un quinto parametro P5 che tenga conto di questa caratteristica. Per ogni combinazione, si calcola la percentuale di geni che presenta un profilo di espressione crescente. In corrispondenza del valore percentuale ottenuto, viene assegnato al parametro un valore compreso tra 0 e 1.

La procedura di assegnazione è molto simile a quella illustrata per il quarto parametro: ad un valore percentuale molto basso o molto alto corrisponde un valore del parametro tendente a 1, invece un valore basso del parametro, tendente a 0, scaturisce da un valore percentuale prossimo al 50%.

Il sesto parametro P6 riguarda l'individuazione, all'interno del gruppo, di sequenze geniche con un'escursione massima nello stesso istante temporale.

Un agente esterno, come ad esempio la somministrazione di una determinata sostanza o la variazione delle condizioni ambientali, quale ad esempio un innalzamento o un abbassamento della temperatura, potrebbe provocare un forte aumento o una forte diminuzione del livello di espressione genica. La presenza di picchi, in corrispondenza dello stesso istante temporale, può portare all'individuazione di un gruppo di geni che risponda in maniera simile in presenza di un agente esterno.

Tenendo conto che i valori di espressione genica sono stati normalizzati e quindi variano tra 0 e 1, si considera un valore di soglia pari a metà dell'ampiezza dell'intervallo di normalizzazione (0.5).

Per ogni sequenza, appartenente alla combinazione da esaminare, si calcola il valore di escursione massima. Se nessun gene della combinazione ha un'escursione massima che supera la soglia, viene assegnato un valore nullo all'ultimo parametro P6. Se tutti i geni del gruppo presentano un "*picco*" (cioè una variazione massima maggiore della soglia) in corrispondenza dello stesso istante temporale, al parametro è assegnato un valore unitario. In tutti gli altri casi al parametro è assegnato un valore che è pari alla percentuale di geni che presenta un *picco* nello stesso istante.

Sotto-sistema intelligente

Il sotto-sistema intelligente proposto è basato su metodologie di Soft Computing, preferibilmente si tratta di un sistema neuro-fuzzy in cui le regole

1. possono essere introdotte dall'utente in maniera linguistica attraverso clausole

del tipo IF..THEN;

2. possono essere generate mediante l'ausilio di una rete neurale con pesi e soglie rappresentanti i parametri caratteristici.

Relativamente al secondo caso, il sotto-sistema deve essere preventivamente addestrato (off-line learning) con un opportuno insieme di dati (matrice di learning), come quello esemplificato in Figura 8.

Nel funzionamento on-line, l'uscita del sistema Fuzzy (valore caratteristico) è confrontata con un valore di soglia THRESHOLD. Tra i gruppi di geni ottenuti, quelli a cui è associato un valore caratteristico superiore alla soglia THRESHOLD vengono identificati come Gene Network, mentre gli altri gruppi vengono scartati.

ESEMPIO APPLICATIVO

Al fine di rendere più chiaro il metodo dell'invenzione, se ne mostra un esempio di applicazione. I dati in ingresso sono costituiti dai livelli di espressione genica di opportune sequenze da esaminare. La Tabella 2 mostra una porzione del set di dati utilizzati per l'esperimento proposto.

Nella prima colonna della Tabella 2 sono riportati gli accession number dei geni, che in questo caso specifico appartengono al genoma del lievito *S. cerevisiae* (si osservi che la prima lettera di ogni accession number è Y che sta per "yeast", che in inglese vuol dire "lievito"). Per ogni gene sono stati estratti i livelli di espressione genica temporalmente (il nome dell'esperimento è ALPHA caratterizzato da 18 istanti temporali) alla fine di un ciclo di divisione cellulare dopo la sincronizzazione mediante il *fattore di arresto alpha*. Tutte le misure sono state effettuate prendendo il valore di espressione genica all'istante t=0 come campione di riferimento (seconda colonna della tabella); le altre colonne, come si osserva dalla Tabella 2, si riferiscono ai livelli di espressione genica misurati rispettivamente dopo 7 min, 14 min e così via dall'esperimento.

Per ciascun gene, identificato dal corrispondente accession number, è possibile avere delle informazioni aggiuntive, quali ad esempio la descrizione

(Description), la categoria funzionale (Molecular Function) e le *annotation* (Biological Process) del gene specifico. Queste informazioni sono disponibili nel *Saccharomyces* Genome Database. Un esempio è mostrato nella Tabella 3.

Risultati ottenuti con un criterio di raggruppamento

Sono state considerate 1533 sequenze geniche relative al genoma del lievito. Ciascuna sequenza è costituita da diciotto livelli di espressione genica, corrispondenti ai valori misurati nel tempo, ad intervalli di 7 minuti, dopo l'esperimento ALPHA (istante t=0).

Queste sequenze sono state raggruppate mediante l'algoritmo K-means adottando un numero di centroidi iniziale pari a 50 e un metodo random di generazione dei centroidi.

Alla fine del processo di clustering sono state ottenute 50 sotto-tabelle (CLUSTERS) pari al numero di centroidi scelto inizialmente nella fase di selezione dei parametri. Questa condizione indica il fatto che non si è avuta la formazione di CLUSTER vuoti, che eventualmente sarebbero stati scartati alla fine del processo di raggruppamento.

Nella Tabella 4 è mostrato il contenuto del file *kmyeast50.txt* relativo al cinquantesimo CLUSTER costituito da 9 sequenze geniche. Nel file sono indicati gli accession number (GenBank) e i diciotto valori di espressione genica delle sequenze (PARAMETERS) costituenti il CLUSTER.

Fase di Filtraggio

Secondo l'invenzione si può opzionalmente effettuare una fase di filtraggio mediante la quale è possibile selezionare alcune delle sequenze geniche in esame in funzione dei valori di espressione assunti nei diversi istanti temporali. Ad esempio si possono filtrare tutte le sequenze che all'istante t=0 hanno un valore di espressione genica maggiore di zero oppure si può considerare un criterio di filtraggio relativo a più parametri contemporaneamente. Tuttavia questa fase, come già detto, è opzionale e quindi per semplicità nell'esempio riportato non è

stata considerata.

Generazione di combinazioni

In questa fase vengono generate tutte le combinazioni di tipo CLUSTER-CLUSTER. È evidente che nel caso in cui fosse stata eseguita la fase di filtraggio sarebbero state generate anche le combinazioni di tipo FILTER-FILTER e CLUSTER-FILTER.

In questo caso, considerando che il numero di combinazioni è dato da $\binom{K}{2} = \frac{K!}{(K-2)!*2!} = \frac{K*(K-1)}{2}$ e tenendo conto che il numero K di CLUSTER generato nella fase precedente è pari a 50, si è ottenuto un numero di combinazioni pari a 1225.

Per ciascuna combinazione sono stati estratti i sei parametri di natura numerica P1, ..., P6 introdotti precedentemente. In funzione dei parametri estratti il sotto-sistema intelligente ha assegnato a ciascuna combinazione un valore caratteristico compreso tra 0 e 1. Tutte le combinazioni con un valore caratteristico (*grado di Gene Network*), maggiore di una predeterminata soglia sono state individuate come possibili Gene Network. In questo esempio è stato considerato un valore di soglia pari a 0.5 e sono stati individuati sei possibili Gene Networks.

Nel file *gnyeast.txt* mostrato in Tabella 5 sono indicate tutte le informazioni riguardanti le combinazioni generate.

Nella prima colonna sono indicati i nomi dei files contenenti informazioni più dettagliate sulle combinazioni generate. I nomi dei files che iniziano con la lettera X si riferiscono alle combinazioni a cui il sotto-sistema intelligente ha assegnato un grado minore di 0.5; i rimanenti files, invece, si riferiscono alle combinazioni che il sistema ha identificato come possibili gene network e che in questo caso specifico sono 6.

Nella seconda colonna è riportato il numero di sequenze geniche costituente la combinazione in esame, mentre nelle rimanenti colonne è indicato il tipo di

combinazione (esempio CLUSTER22-CLUSTER26).

L'ultima colonna rappresenta il valore assegnato dal sistema neuro-fuzzy precedentemente addestrato. È evidente che quanto più il grado assegnato è prossimo all'unità, tanto più la combinazione in esame si presta meglio ad essere una possibile rete regolatoria. Viceversa, quanto più il grado d'uscita è prossimo a zero, tanto maggiore è l'incertezza sulla possibilità che la combinazione abbia una relazione intrinseca.

Ad esempio (terza riga), alla combinazione tra il CLUSTER26 (C26), costituito da 9 sequenze geniche, e il CLUSTER30 (C30), costituito da 21 sequenze, è stato assegnato un grado pari a 0.67 e pertanto è stata indicata come possibile gene network. Nelle Tabelle 6 e 7 sono riportate le informazioni relative al CLUSTER26 (C26) e al CLUSTER30 (C30).

Tutte le informazioni necessarie relative al data set combinato sono illustrate per completezza nella Tabella 8.

Nella seconda colonna della Tabella 8 è indicato il CLUSTER di provenienza della sequenza genica il cui accession number è riportato nella prima colonna. Oltre ai valori di espressione genica delle sequenze costituenti la combinazione, sono stati anche riportati i valori dei sei parametri estratti e il relativo grado d'uscita assegnato dal sistema esperto.

Tralasciando il calcolo dei primi tre parametri P1, P2 e P3 relativi alla correlazione lineare e quadratica della combinazione, per maggiore chiarezza viene riportata la procedura relativa al calcolo di P4, P5 e P6 per la combinazione C26-C30 indicata in Tabella 8.

Calcolo di P4

Per il calcolo del parametro P4, percentuale di geni con valore finale maggiore del valore iniziale, si deve considerare, per ogni sequenza della combinazione, la variazione Δ tra il valore di espressione genica corrispondente all'ultimo istante

temporale *alpha*119 e il valore di espressione corrispondente al primo istante *alpha*0. Per la prima sequenza, YPR120C, la variazione Δ è pari a:

$$\Delta = -0.43 - (-0.92) = 0.49 \quad \Rightarrow \quad \Delta > 0$$

Questo calcolo viene effettuato per tutte le sequenze della combinazione. Si ottiene che 21 sequenze su 30, ovvero il 70% delle sequenze ha una variazione positiva (Δ>0). Questa percentuale, che nel seguito sarà indicata con la variabile "VALORE_PERCENTUALE" è poi convertita in un valore compreso tra 0 e 1.

La procedura di conversione è la seguente:

- Se VALORE_PERCENTUALE=50%, a P4 è assegnato un valore nullo.
- Se 50%<VALORE_PERCENTUALE≤70%, a P4 è assegnato un valore compreso tra 0 e 0.5 (più alto è il valore percentuale, più alto è il valore del parametro).
- Se 70%<VALORE_PERCENTUALE≤100%, a P4 è assegnato un valore compreso tra 0.5 e 1 (più alto è il valore percentuale, più alto è il valore del parametro).
- Se 30%≤VALORE_PERCENTUALE<50%, a P4 è assegnato un valore compreso tra 0 e 0.5 (più basso è il valore percentuale, più alto è il valore del parametro).
- Se 0%≤VALORE_PERCENTUALE<30%, a P4 è assegnato un valore compreso tra 0.5 e 1 (più basso è il valore percentuale, più alto è il valore del parametro).

In pratica, più la percentuale di geni aventi la stessa variazione è prossima al 50%, più il valore del quarto parametro P4 si approssima a 0. Più invece VALORE_PERCENTUALE si avvicina al 100%, più il parametro P4 si avvicina a 1, in quanto la maggior parte dei geni del gruppo si comporta allo stesso modo.

Nel caso in cui la percentuale è bassa e prossima a 0, il valore del parametro P4 è elevato e tende a 1. Ciò è dovuto al fatto che percentuali basse di geni aventi una variazione positiva (Δ>0), implicano percentuali elevate di sequenze con variazione negativa (Δ<0). Questo parametro consente di individuare di gruppi di geni con un comportamento simile dal punto di vista della variazione complessiva del valore di espressione genica, prescindendo dal segno della variazione.

Nell'esempio considerato, dato che il valore percentuale è pari a 70%, a P4 viene assegnato un valore pari a 0.5 (si veda Tabella 8).

Calcolo di P5

Per il calcolo del parametro P5, percentuale di geni con lo stesso andamento temporale, si deve verificare, per ogni sequenza della combinazione, se il profilo di espressione è crescente o decrescente. Dato che i valori di espressione genica sono discretizzati nel tempo, bisogna calcolare le variazioni Δi tra il valore di espressione genica corrispondente all'istante (i)-esimo e il valore di espressione corrispondente all'istante (i-1)-esimo per i=1, 2, ..., n e con n uguale al numero di esperimenti. In questo caso specifico n=18 e per ogni sequenza vengono calcolati n-1 (17) valori di variazione.

Ad esempio, per la prima sequenza YPR120C i valori Δi sono i seguenti:

$$\Delta 1 = alpha7\text{-}alpha0 = -0.32 - (0.92) = 0.6 > 0$$
$$\Delta 2 = alpha14\text{-}alpha7 = 0.98 - (-0.32) = 1.3 > 0$$
$$\Delta 3 = alpha21\text{-}alpha14 = 1.03\text{-}0.98 = 0.05 > 0$$
$$\Delta 4 = alpha28\text{-}alpha21 = 0.32\ \text{-}1.03 < 0$$
$$\Delta 5 = alpha35\text{-}alpha28 = -0.03\text{-}0.32 < 0$$
$$\Delta 6 = alpha42\text{-}alpha35 = -0.12\text{+}0.03 < 0$$
$$\Delta 7 = alpha49\text{-}alpha42 = -0.34\text{+}0.12 < 0$$
$$\Delta 8 = alpha56\text{-}alpha49 = -0.29\text{+}0.34 > 0$$
$$\Delta 9 = alpha63\text{-}alpha56 = -0.27\text{+}0.29 > 0$$
$$\Delta 10 = alpha70\text{-}alpha63 = 0.76\text{+}0.27 > 0$$
$$\Delta 11 = alpha77\text{-}alpha70 = 0.67\text{-}0.76 < 0$$
$$\Delta 12 = alpha84\text{-}alpha77 = 0.37\text{-}0.67 < 0$$
$$\Delta 13 = alpha91\text{-}alpha84 = -0.17\text{-}0.37 < 0$$
$$\Delta 14 = alpha98\text{-}alpha91 = 0.16\text{+}0.17 > 0$$
$$\Delta 15 = alpha105\text{-}alpha98 = -0.14\text{-}0.16 < 0$$
$$\Delta 16 = alpha112\text{-}alpha105 = -0.15\ \text{+}0.14 < 0$$
$$\Delta 17 = alpha119\text{-}alpha112 = -0.43\ \text{+}0.15 < 0$$

Se il numero di variazioni Δi positive è maggiore del numero di variazioni Δi negative, la sequenza ha un profilo temporale complessivamente crescente; viceversa la sequenza ha un profilo complessivamente decresente.

Nel caso della sequenza YPR120C, il numero di variazioni Δi positive è 7, mentre il numero di variazioni Δi negative è 10. Dato che il numero di variazioni Δi positive è minore del numero di variazioni Δi negative, alla sequenza viene attribuito un profilo di espressione temporale decrescente.

Lo stesso calcolo viene ripetuto per ognuna delle rimanenti 29 sequenze della combinazione illustrata in Tabella 8, ottenendo che una certa percentuale di sequenze (indicata in seguito con la variabile "PERCENT"), ha un profilo temporale complessivamente crescente. A questo valore percentuale viene associato un parametro P5, avente un valore compreso tra 0 e 1, con una procedura analoga a quella esposta per il parametro P4.

Nel caso considerato, si ha che 5 sequenze su 30, cioè il 16,7% ha un profilo temporale complessivamente crescente, per cui P5 assume un valore alto. Infatti l'83.3% delle sequenze ha un profilo decrescente e quindi una percentuale molto alta di sequenze si comporta in maniera simile dal punto di vista dell'andamento temporale.

Si ribadisce che il valore del parametro P5 non dipende dal fatto che l'andamento complessivo della maggioranza dei geni è crescente o decrescente, ma da quanti geni della combinazione hanno lo stesso andamento complessivo. Nel caso esemplificato, il parametro (P5) è prossimo all'unità.

Una possibile procedura di valutazione di P5 è descritta con maggior dettaglio di seguito.

Si definiscono tre valori di soglia

SOGLIA1=0.3; SOGLIA2=1-SOGLIA1=0.7; SOGLIA3=0.5;

e in funzione dei valori di soglia vengono calcolati i seguenti valori:

VALORE1=((SOGLIA2-SOGLIA3)/(1-SOGLIA3))=0.4;

VALORE2=((2*SOGLIA2-1+SOGLIA3)/(2*SOGLIA3))=0.9;

- Se SOGLIA2<=PERCENT<=1, P5=((PERCENT -VALORE1)/(1-VALORE1));
- Se 0<=PERCENT<=SOGLIA1, P5=(((1-PERCENT)-VALORE1)/(1-VALORE1));
- Se 0.5<=PERCENT<SOGLIA2, P5=((PERCENT-0.5)/(VALORE2-0.5));
- Se SOGLIA1<PERCENT<0.5, P5=(((1-PERCENT)-0.5)/(VALORE2-0.5));

Si osservi che ad una percentuale del 50% corrisponde un valore di P5 nullo perché, come si è già ampiamente detto, in questo caso non si ha un andamento temporale complessivo (crescente o decrescente) predominante da parte delle sequenze costituenti la combinazione.

Nell'esempio proposto (Tabella 8) il valore di P5 è dato da:

P5=(((1-PERCENT)-VALORE1)/(1-VALORE1))=(1-0.167-0.4)/0.6=0.72

Calcolo di P6

Per il calcolo del parametro P6, percentuale di geni con massima escursione nello stesso istante temporale, si verifica, per ogni sequenza della combinazione, se in valore assoluto la variazione Δi supera un determinato valore di soglia. Dato che i valori relativi ad ogni esperimento sono stati normalizzati tra 0 e 1, per il valore di soglia si è scelto 0.5 pari a metà dell'ampiezza dell'intervallo di normalizzazione. In Tabella 9 sono stati riportati i valori di espressione genica normalizzati tra 0 e 1 delle sequenze costituenti la combinazione C26-C30.

Per la sequenza YPR120C si ha che:

$$|\Delta 1| = |alpha7 - alpha0| = 0.307$$

$$|\Delta 2| = |alpha14 - alpha7| = 0.666$$

.....

$$|\Delta 17| = |alpha119 - alpha112| = 0.14359$$

Ripetendo questi calcoli per tutte le sequenze si ottengono i risultati riassunti in

Tabella 10.

I valori $|\Delta|$ che superano la soglia sono sottolineati. Per ogni sequenza il massimo valore $|\Delta|$ che supera la soglia rappresenta il *picco* (escursione massima) ed è racchiuso in un riquadro. Ad esempio, in corrispondenza della prima sequenza YPR120C si hanno due valori maggiori della soglia, $|\Delta 2| = 0.66$ e $|\Delta 10| = 0.52$; il picco in questo caso è rappresentato da $|\Delta 2|$.

Si osservi che non tutte le sequenze necessariamente devono presentare un picco. Nell'esempio proposto le sequenze YJL115W, YCR065W, YOR074C, YKL113C, YKL076W, YER001W e YDR309C non presentano un picco.

Per il calcolo del sesto parametro P6 si deve considerare il numero massimo di picchi in corrispondenza dello stesso istante temporale. In questo esempio, il numero massimo di picchi è 17 e si ha in corrispondenza di $|\Delta 2|$. In particolare il 56.7% (17 sequenze su 30) delle sequenze della combinazione presenta un picco in corrispondenza dello stesso istante temporale e quindi in questo caso P6 sarà uguale a 0.57, come riportato in Tabella 8.

RIFERIMENTI BIBLIOGRAFICI

[1] Luke Alphey, *DNA Sequencing:from experimental methods to bioinformatics*, BIOS Scientific Publishers, 1997.

[2] G. Lewin, *Gene IV*, 1998.

[3] D.L. Kirk, *Biologia Oggi*, Piccinini editore Padova.

[4] Chieffi, Dolfini, Malcovati, Pierantoni, Tenchini, *Biologia e Genetica*, EdiSES.

[5] M.L.M. Anderson, *Nucleic Acid Hybridization*, BIOS scientific Publishers, 1999.

[6] M. Schena, *DNA Microarray: A Practical Approach*, Oxford University Press 1999.

[7] Patrik D'haeseller, *Reconstructing Gene Networks from Large Scale Gene Expression Data*, The University of New Mexico, December 2000.

[8] Pierre Bladi, Soren Brunak, *Bioinformatics. The Machine Learning Approach*, MIT Press 1998.

[9] Primer on Molecular Genetic, *DOE Human Genome Program*, 1992.

[10] M.B.Eisen, P.T.Spellman, Patrick O.Brown, D.Botstein, *Cluster Analysis and display of genome-wide expression patterns*, Vol. 95, pp. 14863-14868, December 98, Genetics.

[11] P.D'haeseleer, S.Liang, R. Somogyi, *Genetic Network inference: from co-expression clustering to reverse engineering*, Vol.16 no.8 2000, pages 707-726.

[12] S.Huang, *Gene expression profiling, genetic networks and cellular states: an integrating concept for tumorigenesis and drug discovery*, 1 July 1999.

[13] P.Smolen, D. A. Baxter, J.H. Byrne, *Mathematical Modeling of Gene Networks*, Neuron, Vol.26, 567-580, June 2000.

[14] Frank Höppner, Frank Klawonn, "*Fuzzy Cluster Analysis: Methods for Classification Data Analysis and Image Recognition*", prima edizione 1999.

[15] Michele Scardi, "*Tecniche di Analisi dei Dati in Ecologia*", Versione 1.2a, Aprile 1998.

[16] Daniel Boley, Vivian Borst, "*Unsupervised Clustering: A Fast Scalable Method for Large Datasets*", Department of Computer Science and Engineering.

[17] D.L. Boley, Principal Direction Divisive Partitioning, *Data Mining and Knowledge Discovery*, 1999, Department of Computer Science and Engineering.

[18] Kohonen T. *Self-Organization and Associative Memory*, Primavera 1984.

[19] Flexer A. *On the Use of Self-Organizing Maps for Clustering and Visualization.*

[20] R.J. Hataway, J.C.Bezdek, Y.Hu, *Generalized Fuzzy c-Means Clustering Strategies Using Lp Norm Distances*, IEEE Transactions on fuzzy system, Vol. 8, No 5, October 2000.

[21] R.Krishapuram, J.Keller, *A Possibilistic Approach to Clustering*, IEEE Transactions on fuzzy system, Vol. 1, No 2, May 1993.

[22] Lofti A. Zadeh, "*Fuzzy logic, Neural Networks, and SoftComputing*", Comm. of the ACM, March 1994, vol 37, No.3.

[23] X L. Daboni, *Calcolo delle probabilità ed elementi di Statistica.*

RIVENDICAZIONI

1. Metodo di analisi di una tabella di dati relativi all'espressione di geni variabile nel tempo o relativa a condizioni differenti, al fine di identificare gruppi di geni co-espressi e co-regolati, comprendente

definire un criterio di raggruppamento (clustering) di dati di detta tabella;

per detto criterio di raggruppamento (clustering), determinare gruppi di geni in sotto-tabelle (CLUSTERS) che soddisfano tale criterio di raggruppamento (clustering);

generare combinazioni di coppie di dette sotto-tabelle;

calcolare parametri caratteristici dei dati associati a geni di una stessa combinazione;

generare un valore caratteristico definito in funzione di detti parametri per ciascuno di detti gruppi di geni mediante un algoritmo di decisione basato su tecniche di Soft Computing;

identificare le combinazioni il cui valore caratteristico è superiore ad una certa soglia prestabilita come '*Gene Networks*' e scartare gruppi di geni il cui valore caratteristico è inferiore a detta soglia.

2. Il metodo della rivendicazione 1, comprendente inoltre le operazioni di

definire un rispettivo insieme di criteri logici di filtraggio di dati di detta tabella;

per ciascun criterio logico, determinare una corrispondente sotto-tabella (FILTER) filtrata contenente dati dei geni i cui valori di espressione soddisfano tale criterio logico;

generare combinazioni di coppie di sotto-tabelle ottenute con detti criteri logici e di raggruppamento (filtering, clustering).

3. Il metodo della rivendicazione 1, in cui detto algoritmo di decisione è un algoritmo fuzzy i cui antecedenti e conseguenti sono definiti in funzione di detti parametri.

4. Il metodo della rivendicazione 1, in cui detti parametri sono scelti

nell'insieme composto da parametri numerici legati al profilo di espressione genica, parametri che hanno un significato biologico di natura semantica, e da parametri misti che esprimono contemporaneamente relazioni di natura numerica e di natura semantica.

5. Il metodo della rivendicazione 1, in cui detti parametri ed indici di correlazione sono scelti nell'insieme costituito da:

valori assoluti dei coefficienti di correlazione lineare tra dati associati a coppie di geni;

valori assoluti dei coefficienti di correlazione quadratica tra dati associati a coppie di geni;

percentuale di geni della combinazione che ha valore di espressione genica finale maggiore del rispettivo valore di espressione genica iniziale;

percentuale di geni della combinazione che ha valore di espressione genica finale minore del rispettivo valore di espressione genica iniziale;

percentuale di geni i cui valori di espressione genica hanno uno stesso andamento temporale crescente o decrescente;

percentuale di geni che presenta un massimo valore di espressione genica in una stessa condizione;

percentuale di geni che hanno attributi (ontologies) in comune;

percentuale di geni che hanno domini funzionali in comune.

6. Il metodo della rivendicazione 1, comprendente inoltre scartare combinazioni tra sotto-tabelle composte da un numero di geni inferiore ad un certo numero prestabilito, introducendo solamente una volta geni che sono compresi in entrambe le sotto-tabelle combinate.

7. Il metodo della rivendicazione 1, in cui detti criteri di raggruppamento (clustering) sono basati su algoritmi scelti nell'insieme composto dagli algoritmi gerarchico agglomerativo, non gerarchico Kmeans, gerarchico Kmeans sequenziale, non gerarchico SOM e non esclusivo Fuzzy Clustering.

8. Il metodo della rivendicazione 5, comprendente

calcolare i coefficienti di correlazione tra tutte le sequenze geniche della combinazione;

suddividere l'intervallo di valori da 0 a 1 in cinque sotto-intervalli di uguale ampiezza e assegnare a ciascuno di detti sotto-intervalli un rispettivo valore di correlazione quantizzato (v_i);

calcolare la percentuale di coefficienti di correlazione appartenenti a ciascun sotto-intervallo;

definire per ciascuna combinazione un coefficiente complessivo di correlazione lineare ottenuto come media aritmetica dei valori quantizzati associati ai sotto-intervalli in cui è distribuito un numero di coefficienti maggiore del 50%.

9. Il metodo della rivendicazione 5, comprendente

calcolare coefficienti di correlazione quadratica tra tutte le sequenze geniche di una stessa combinazione;

definire per ciascuna combinazione un coefficiente complessivo di correlazione quadratica ottenuto come media aritmetica di detti valori di correlazione.

10. Il metodo della rivendicazione 5, comprendente

calcolare la percentuale di sequenze della combinazione con un valore di espressione genica finale maggiore di quella iniziale;

definire un coefficiente relativo alla variazione complessiva del valore di espressione genica compreso tra 0 e 1 corrispondente a detta percentuale;

11. Il metodo della rivendicazione 5, comprendente

calcolare la percentuale di sequenze della combinazione con un andamento temporale crescente;

definire un coefficiente relativo all'andamento temporale del profilo di espressione genica compreso tra 0 e 1 corrispondente a detta percentuale.

12. Il metodo della rivendicazione 5, comprendente

calcolare la percentuale di sequenze della combinazione con un valore di espressione maggiore di una prefissata soglia in corrispondenza di uno stesso istante;

definire un coefficiente relativo alla presenza di escursioni massime del livello di espressione genica in corrispondenza dello stesso istante temporale compreso tra 0 e 1 corrispondente a detta percentuale.

13. Sistema di identificazione di gruppi di geni co-espressi e co-regolati secondo il metodo della rivendicazione 1, comprendente

un sotto-sistema di pre-elaborazione (pre-processing), ricevente in ingresso dati di una tabella relativi all'espressione di geni variabile nel tempo o relativa a condizioni differenti, generante sotto-tabelle (CLUSTERS) di dati di gruppi di geni che soddisfano un criterio di raggruppamento predefinito;

un sotto-sistema di elaborazione dei dati di dette sotto-tabelle (CLUSTERS), generante segnali rappresentativi di parametri caratteristici dei dati associati a geni di una stessa combinazione di coppie di dette sotto-tabelle;

un sotto-sistema intelligente ricevente in ingresso detti segnali rappresentativi di parametri caratteristici e producente in uscita dati di gruppi di geni identificati come 'Gene Networks'.

14. Il sistema della rivendicazione 13, in cui detto sotto-sistema intelligente è un sotto-sistema fuzzy addestrato off-line identificato mediante una rete neuronale.

ORF	0 Minutes	30 Minutes	1 Hour	2 Hours
YAL001C	1	1.3	2.4	5.8
YAL002W	0.9	0.8	0.7	0.5
YAL003W	0.8	2.1	4.2	10.1
YAL005C	1.1	1.3	0.8	
YAL010C	1.2	1	1.1	4.5

TAB. 1

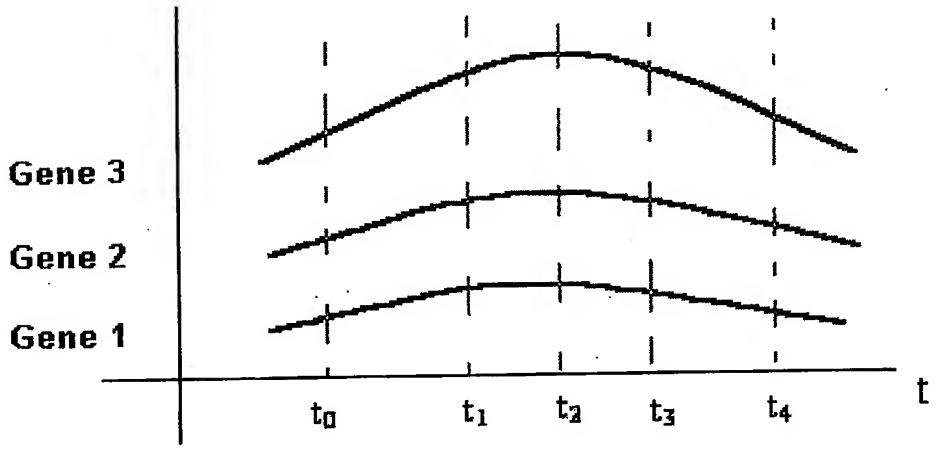

FIG. 1

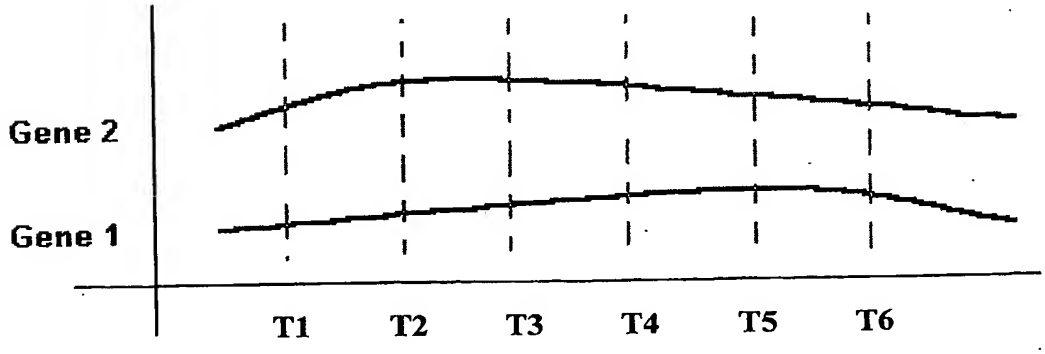

FIG. 2

NCBI Reference Sequences (RefSeq)

Category: PROVISIONAL

mRNA:	NM_017106		
Protein:	NP_058802	chloride channel 5	**BL**
Domains:	CBS domain	score: 103	
	Biotin synthase	score: 86	
	Voltage gated chloride channels	score: 858	
	Domain in cystathionine beta-synthase and other proteins.	score: 108	
GenBank Source:	D50497		

GenBank Sequences

Nucleotide	Type	Protein	
D50497	m	BAA09091	**BL**
Z56277	m	CAA91216	**BL**

Function Submit GeneRIF

Gene Ontology™:

Term	Evidence	Source	Pub
• membrane	IEA	RGD	
• chloride transport	IEA	RGD	

FIG. 3

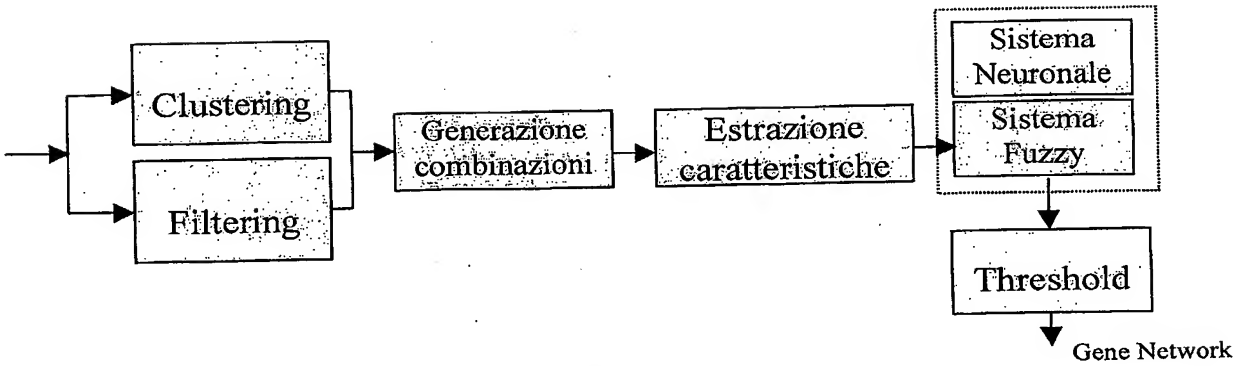

FIG. 4

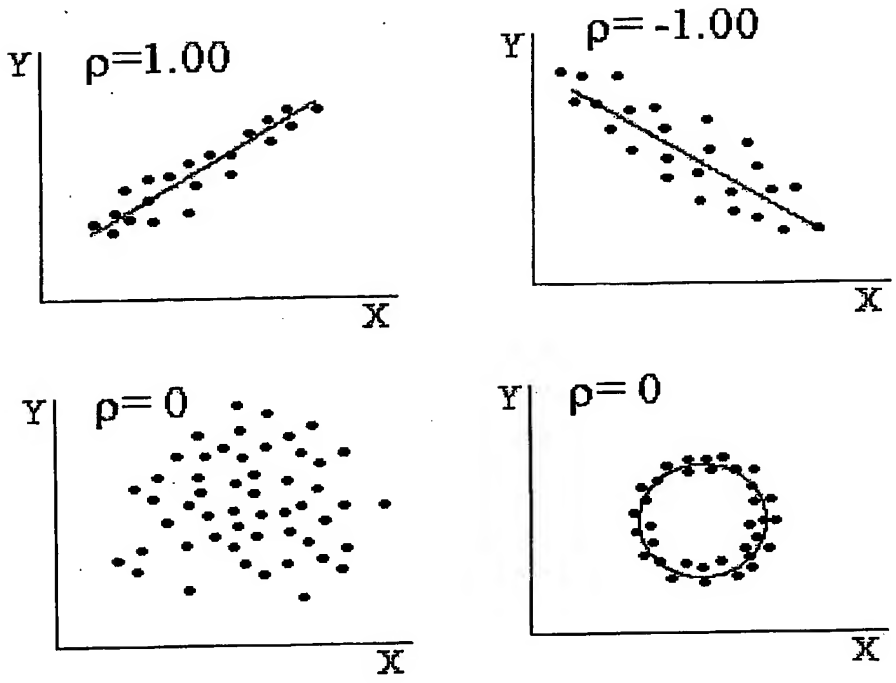

FIG. 5

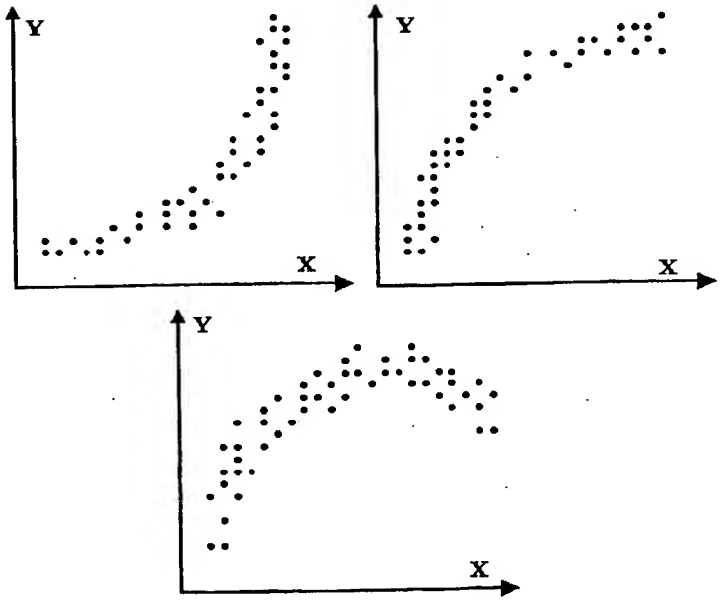

FIG. 6

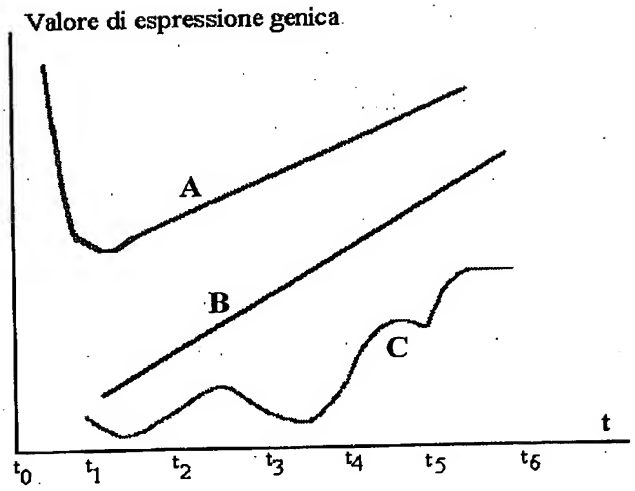

FIG. 7

P1	P2	P3	P4	P5	P6	Out
0.90	0.90	1.00	0.83	1.00	0.90	1.00
0.10	0.10	1.00	1.00	0.47	0.28	0.00
0.90	0.90	1.00	1.00	1.00	0.53	0.00
0.90	0.60	1.00	0.69	1.00	0.75	1.00
0.00	0.00	0.00	0.92	0.92	0.05	0.00
0.90	0.00	1.00	1.00	1.00	1.00	1.00
0.58	0.44	1.00	0.67	0.67	0.60	1.00
0.70	0.43	1.00	0.42	1.00	0.67	0.00
.....	.....	.....	.....	.....	.....	.....
0.84	0.47	1.00	0.00	1.00	0.63	0.00
0.00	0.00	0.00	1.00	0.92	0.05	0.00

FIG. 8

Genbank	alpha 0	alpha 7	alpha 14	.....	Alpha 21	alpha 28	alpha 35
YBR166C	0.33	-0.17	0.04	.....	-0.07	-0.09	-0.12
YOR357C	-0.64	-0.38	-0.32	.....	-0.29	-0.22	-0.01
YLR292C	-0.23	0.19	-0.36	.....	0.14	-0.4	0.16
YDL120W	0.11	0.32	0.03	.....	0.32	0.03	-0.12
YGL248W	-0.25	0.26	0.01	.....	-0.06	-0.42	-0.07
YIL146C	-0.58	-0.29	-0.45	.....	-0.15	-0.86	-0.36
YJR106W	-0.36	-0.17	-0.22	.....	-0.34	-0.36	0.03
YBR123C	-0.17	-0.32	-0.34	.....	-0.42	-0.25	-0.3
.....	.....	.....	.....	.....	.....	.....	.....
YHR047C	-0.29	-0.07	-0.34	.....	-0.34	-0.36	-0.43
YMR055C	-0.34	0.88	-0.42	.....	-0.97	-0.15	-0.29
YDR457W	0.01	-0.69	-0.09	.....	-0.09	0.25	0.21

Tab. 2

GenBank	Description	Molecular Function	Biological Process
YBR166C	TYR1 TYROSINE BIOSYNTHESIS...	GO:8977 prephenate..	GO:6570 tyrosine metabolism
YOR357C	GRD19 SECRETION GOLGI....		GO:8104 protein localization
YLR292C	SEC72 SECRETION ER PROTEIN...	GO:5047 signal recognition...	GO:6615 SRP-dependent...
YDL120W	YFH1 IRON HOMEOSTASIS...	GO:5554 unknown	GO:6879 iron homeostasis
YGL248W	PDE1 PURINE METABOLISM 3',5'...	GO:4115 cAMP-specific...	GO:19933 cAMP-mediated..
YIL146C	ECM37 CELL WALL BIOGENESIS...	GO:5554 unknown	GO:7047 cell wall organization..
YJR106W	ECM27 CELL WALL BIOGENESIS...	GO:5554 unknown	GO:7047 cell wall organization...
YBR123C	TFC1 TRANSCRIPTION...	GO:3709 RNA polymerase III...	GO:6384 transcription initiation...
.....	.....	.....	.....

Tab. 3

NUMERO DI GENENETWORK:	6					
NUMERO MAX DI ELEMENTI:	1533					GRADO
gnyeast1.txt	10	kmyeast22.txt	kmyeast26.txt	CLUSTER 22	CLUSTER 26	1
gnyeast2.txt	19	kmyeast22.txt	kmyeast34.txt	CLUSTER 22	CLUSTER 34	0.91
gnyeast3.txt	30	kmyeast26.txt	kmyeast30.txt	CLUSTER 26	CLUSTER 30	0.67
gnyeast4.txt	13	kmyeast26.txt	kmyeast44.txt	CLUSTER 26	CLUSTER 44	0.68
gnyeast5.txt	10	kmyeast26.txt	kmyeast45.txt	CLUSTER 26	CLUSTER 45	1
gnyeast6.txt	19	kmyeast34.txt	kmyeast45.txt	CLUSTER 34	CLUSTER 45	0.8
Xgnyeast1.txt	34	kmyeast1.txt	kmyeast2.txt	CLUSTER 1	CLUSTER 2	0
Xgnyeast2.txt	43	kmyeast1.txt	kmyeast3.txt	CLUSTER 1	CLUSTER 3	0.3
Xgnyeast3.txt	57	kmyeast1.txt	kmyeast4.txt	CLUSTER 1	CLUSTER 4	0
Xgnyeast4.txt	38	kmyeast1.txt	kmyeast5.txt	CLUSTER 1	CLUSTER 5	0
Xgnyeast5.txt	54	kmyeast1.txt	kmyeast6.txt	CLUSTER 1	CLUSTER 6	0
Xgnyeast6.txt	59	kmyeast1.txt	kmyeast7.txt	CLUSTER 1	CLUSTER 7	0
.....	.....	.....	.....	.....	.....	.....

Tab. 5

Cluster N. 50																			
- Genebank -	- PARAMETERS -																		Description
YPL111W	0.16	0.1	0.86	1.01	0.99	1.23	1.9	1.41	1.51	0.88	1.04	1.06	1.04	0.44	0.72	0.75	0.95	0.88	CAR1 ARGININE METABOLISM ARGIN.
YBL005W	1.21	0.82	1.23	1.34	1.14	1.14	1	0.3	1.24	1.23	0.72	1.06	0.64	1.09	0.91	1.12	0.89	1.22	PDR3 TRANSPORT TRANSCRIPT
YJR028W	1.34	1.28	1.29	0.77	1.59	1.1	1	1.46	1.22	1.2	1.79	1	0.63	1.37	0.88	0.96	0.38	0.86	NONE TRANSCRIPTION TFIIE 66 KD
YGR112W	1.23	1.02	1.01	1.19	1.04	1.21	0.7	1.32	0.82	0.57	1.43	0.93	0.53	0.89	1	1.24	0.69	0.84	SHY1 RESPIRATION MITOCHONDF
YMR058W	0.82	-0.15	0.04	0.16	0.74	0.82	1.5	1.17	1.69	1.65	1.96	1.7	1.56	1.21	1.92	1.74	2.11	1.65	FET3 TRANSPORT CELL SURFAC
YOL058W	0.55	1.66	1.94	1.58	1.3	0.9	1.1	0.65	0.7	0.55	0.93	0.62	0.9	0.87	1.24	0.8	1.14	1.07	ARG1 ARGININE BIOSYNTHESIS ARGINI
YMR011W	1.1	1.84	2.12	1.65	1.1	1	0.7	0.08	0.36	0.85	1.99	1.83	1.55	0.7	0.62	-0.23	-0.04	-0.09	HXT2 TRANSPORT HEXOSE PERI
YNL036W	0.01	3.1	2.97	2.83	2.19	1.9	1.8	1.04	1.21	0.93	0.91	0.42	0.76	0.3	0.73	0.67	0.76	0.93	NCE103 SECRETION, NON-CLASSICAL UN
YHR071W	0.77	1.48	1.96	2.16	1.61	1.38	1.3	1.17	1.18	0.54	0.88	0.65	0.96	0.78	0.99	0.97	0.37	0.67	PCL5 CELL CYCLE CYCLIN (PHO8

Tab. 4

Cluster N.26																			
- Genebank -	PARAMETERS																		Description
YDR097C	-0.56	-0.69	0.7	1.2	1	0.4	-0.5	-0.67	-0.2	-0.54	1.16	1.24	0.81	0.34	0.11	0.19	-0.6	-0.49	MSH6 DNA REPAIR M
YOR074C	-1.43	-0.6	0.28	0.79	0.88	0.28	0.01	-1.03	-0.97	-0.4	-0.67	0.45	0.44	-0.2	-0.56	-0.51	-0.92	-1.09	CDC21 DNA REPLICATION
YER070W	-1.22	-0.51	1.32	1.74	0.99	0.71	-0.5	-0.43	-0.79	-0.3	0.59	1.49	0.97	0.44	0.24	0.36	-0.29	-0.47	RNR1 DNA REPLICATION
YBR088C	-1.47	-1.18	0.89	1.29	0.8	-0.17	-0.8	0.48	-1.56	-0.94	0.3	0.97	0.76	-0.06	-0.29	-0.84	-1.12	-1.22	POL30 DNA REPLICATION
YER001W	-2.18	-0.58	0.87	1.71	0.64	0.66	-0.3	-0.43	-0.97	-0.84	0.18	1.46	1.13	1.1	0.31	0.07	-0.86	-0.76	MNN1 PROTEIN GLYCOSYLA
YOL007C	-1.43	-1.25	0.83	0.73	0.77	-0.47	-0.3	-1.18	-1.47	-0.71	-0.32	0.58	0.78	0.39	-0.27	-0.4	-0.84	-1.03	CSI2 CELL WALL BIOGENESIS
YPL256C	-1.69	-0.97	1.11	1.69	0.45	-0.07	-0.6	-1.6	-1.79	-1.36	0.07	1.29	0.82	0.28	-0.1	-0.6	-0.67	-1.32	CLN2 CELL CYCLE G1/S
YIL140W	-1.43	-1.03	1.37	0.74	0.26	-0.17	-0.8	-1.18	-1.09	-1.03	-0.45	0.7	0.29	-0.36	-0.32	-0.51	-0.6	-1.32	SRO4 BUD SITE SELECTION,
YMR199W	-1.6	-0.97	1.25	0.83	0.9	0.44	0.03	-0.58	-1.15	-0.81	0.62	1.1	0.95	0.26	0.31	-0.06	-0.45	-0.92	CLN1 CELL CYCLE G1/S

Tab. 6

Cluster N.30																			
- Genebank -	PARAMETERS																		Description
YPR120C	-0.92	-0.32	0.98	1.03	0.32	-0.03	-0.12	-0.34	-0.3	-0.27	0.76	0.67	0.37	-0.17	0.16	-0.14	-0.15	-0.43	CLB5 CELL CYCLE G1
YJL115W	-0.32	0.49	0.61	1.43	0.58	0.3	-0.45	-0.42	-0.1	0.06	0.34	0.58	0.36	-0.1	-0.32	-0.14	-0.42	-0.43	ASF1 TRANSCRIPTION
YCR065W	-1.22	-0.23	0.54	0.66	0.18	0.07	-0.69	-0.47	-0.4	-0.6	0.18	0.77	0.66	0.38	0.1	0.28	-0.4	-0.38	HCM1 TRANSCRIPTION (PU
YKL045W	-1.03	-0.22	0.63	0.61	0.29	-0.09	-0.62	-0.86	-1	0.19	0.65	0.53	0.24	-0.49	-0.32	-0.45	-0.64	-1.43	PRI2 DNA REPLICATION
YNL262W	0.84	-0.51	0.49	0.58	0.87	0.24	-0.18	-0.64	-0.4	-0.49	0.03	0.32	0.43	0.08	0.04	-0.56	-0.32	-0.71	POL2 DNA REPLICATION
YLR103C	-0.64	-0.2	0.9	0.74	0.48	0.07	-0.3	-0.34	-0.5	-0.34	0.4	0.58	0.33	-0.15	-0.25	-0.15	-0.45	-0.38	CDC45 DNA REPLICATION
YNL312W	-0.69	-0.79	0.48	0.96	0.78	0.77	0.04	-0.47	-0.8	-0.56	0.06	0.23	0.53	-0.15	0.06	-0.62	-0.22	-0.54	RFA2 DNA REPAIR R
YJL074C	-0.74	-1.06	0.46	1.06	0.89	0.04	-0.15	-0.79	-0.8	-0.3	0.12	0.64	0.63	-0.17	-0.27	-0.45	-0.43	-0.2	SMC3 CHROMATIN STRUC
YJL187C	-0.94	-0.64	-0	0.51	0.38	-0.12	-0.2	-0.25	-0.5	-0.74	0.23	0.59	0.58	0.2	0.29	0.14	-1.94	-0.49	SWE1 CELL CYCLE NE
YNL102W	-0.62	2.13	0.19	0.99	0.62	-0.17	-0.22	-0.2	-0.1	-0.64	0.28	0.73	0.71	0.08	0.2	-0.54	-0.69	-0.47	POL1 DNA REPLICATION
YKL113C	-1.12	-0.45	0.29	0.79	0.3	-0.04	-0.56	-0.79	-0.9	-0.71	0.24	0.55	0.5	-0.27	-0.18	-0.25	-0.89	-0.56	RAD27 DNA REPAIR S
YDL164C	-0.62	-0.54	0.55	0.93	0.57	-0.06	-0.1	-0.84	-0.8	-0.4	0.11	0.73	0.6	-0.2	-0.25	-0.6	-0.56	-0.6	CDC9 DNA REPLICATION
YGL038C	-0.86	-0.22	0.5	0.57	-0.4	0.06	-0.69	-0.43	-0.4	0.2	0.32	0.63	0.31	0.14	-0.1	-0.12	-0.45	-0.32	OCH1 PROTEIN GLYCOSYLA
YPL057C	0.32	-0.29	0.96	0.84	0.8	1.08	0.29	-0.45	-0.7	0.19	0.95	0.76	0.58	0.2	0.34	-0.25	-0.42	-0.51	SUR1 SPHINGOLIPID METAE
YKL067W	-0.51	0.21	0.45	1.03	0.77	0.93	0.29	-0.12	-0.4	-0.3	-0.3	-0.03	0.37	-0.14	0.16	-0.23	-0.25	-0.74	YNK1 NUCLEOTIDE METABO
YPR135W	-0.56	-0.76	0.63	1.12	0.51	-0.12	-0.45	-0.79	-0.8	-0.84	0.12	0.57	0.43	-0.29	-0.17	-0.45	-0.42	-0.71	CTF4 DNA REPLICATION
YDR309C	0.53	-0.62	0.33	0.38	0.11	-0.74	-1.09	-1.06	-0.5	-0.3	1.52	0.59	0.64	-0.3	0.53	-0.17	-0.79	-0.42	GIC2 BUD EMERGENCE
YGR152C	-0.49	-0.58	0.8	0.84	0.57	0.34	-0.01	-0.42	-0.5	-0.38	0.43	0.55	0.42	0.21	0.04	-0.3	-0.17	-0.71	RSR1 BUD SITE SELECTION
YBL035C	-0.45	-0.64	1.01	1.14	0.45	-0.4	-0.64	0.15	-1.1	0.44	0.04	0.28	0.32	0.03	-0.54	-0.12	-0.6	-0.3	POL12 DNA REPLICATION
YPR175W	-0.54	-0.69	1.03	0.57	0.49	-0.12	-0.34	-0.62	-0.6	-0.45	0.1	0.52	0.3	-0.22	-0.15	-0.62	-0.2	-0.69	DPB2 DNA REPLICATION
YER111C	-1.25	-0.3	1.32	1.33	0.5	0.14	-0.89	-0.86	-0.8	0.03	0.85	0.74	0.33	-0.23	-0.15	-0.58	-0.38	-0.51	SWI4 CELL CYCLE TR

Tab. 7

Tab. 8

GenBank		alpha 0	alpha 7	alpha 14	alpha 21	alpha 28	alpha 35	alpha 42	alpha 49	alpha 56	alpha 63	alpha 70	alpha 77	alpha 84	alpha 91	alpha 98	alpha 10	alpha 11	alpha 119
YPR120C	C30	-0.92	-0.32	0.98	1.03	0.32	-0.03	-0.12	-0.34	-0.29	-0.27	0.76	0.67	0.37	-0.17	0.16	-0.14	-0.15	-0.43
YJL115W	C30	-0.32	0.49	0.61	1.43	0.58	0.3	-0.45	-0.42	-0.06	0.06	0.34	0.58	0.36	-0.1	-0.32	-0.14	-0.42	-0.43
YCR065W	C30	-1.22	-0.23	0.54	0.66	0.18	0.07	-0.69	-0.47	-0.43	-0.6	0.18	0.77	0.66	0.38	0.1	0.28	-0.4	-0.38
YDR097C	C26	-0.56	-0.69	0.7	1.2	1	0.4	-0.47	-0.67	-0.2	-0.54	1.16	1.24	0.91	0.34	0.11	0.19	-0.6	-0.49
YKL045W	C30	-1.03	-0.22	0.63	0.61	0.29	-0.09	-0.62	-0.86	-1.03	0.19	0.65	0.53	0.24	-0.49	-0.32	-0.45	-0.64	-1.43
YNL262W	C30	0.84	-0.51	0.49	0.58	0.97	0.24	-0.18	-0.64	-0.43	-0.49	0.03	0.32	0.43	0.08	0.04	-0.56	-0.32	-0.71
YOR074C	C26	-1.43	-0.6	0.28	0.79	0.88	0.28	0.01	-1.03	-0.97	-0.4	-0.67	0.45	0.44	-0.2	-0.56	-0.51	-0.92	-1.09
YER070W	C26	-1.22	-0.51	1.32	1.74	0.99	0.71	-0.45	-0.43	-0.79	-0.3	0.59	1.49	0.97	0.44	0.24	0.36	-0.29	-0.47
YLR103C	C30	-0.64	-0.2	0.9	0.74	0.48	0.07	-0.3	-0.34	-0.47	-0.34	0.4	0.58	0.33	-0.15	-0.25	-0.15	-0.45	-0.38
YNL312W	C30	-0.69	-0.79	0.48	0.96	0.78	0.77	0.04	-0.47	-0.79	-0.56	0.06	0.23	0.53	-0.15	0.06	-0.62	-0.22	-0.54
YJL074C	C30	-0.74	-1.06	0.46	1.06	0.89	0.04	-0.15	-0.79	-0.76	-0.3	0.12	0.64	0.63	-0.17	-0.27	-0.45	-0.43	-0.2
YJL187C	C30	-0.94	-0.64	-0.04	0.51	0.38	-0.12	-0.2	-0.25	-0.45	-0.74	0.23	0.59	0.58	0.2	0.29	0.14	-1.94	-0.49
YBR088C	C26	-1.47	-1.18	0.89	1.29	0.8	-0.17	-0.76	0.48	-1.56	-0.94	0.3	0.97	0.76	-0.06	-0.29	-0.84	-1.12	-1.22
YNL102W	C30	-0.62	2.13	0.19	0.99	0.62	-0.17	-0.22	-0.2	-0.09	-0.64	0.28	0.73	0.71	0.08	0.2	-0.54	-0.69	-0.47
YKL113C	C30	-1.12	-0.45	0.29	0.79	0.3	-0.04	-0.56	-0.79	-0.86	-0.71	0.24	0.55	0.5	-0.27	-0.18	-0.25	-0.89	-0.56
YDL164C	C30	-0.62	-0.54	0.55	0.93	0.57	-0.06	-0.1	-0.84	-0.84	-0.4	0.11	0.73	0.6	-0.2	-0.25	-0.6	-0.56	-0.6
YGL038C	C30	-0.86	-0.22	0.5	0.57	-0.36	0.06	-0.69	-0.43	-0.42	0.2	0.32	0.63	0.31	0.14	-0.1	-0.12	-0.45	-0.32
YPL057C	C30	0.32	-0.29	0.96	0.84	0.8	1.08	0.29	-0.45	-0.74	0.19	0.95	0.76	0.58	0.2	0.34	-0.25	-0.42	-0.51
YKL067W	C30	-0.51	0.21	0.45	1.03	0.77	0.93	0.29	-0.12	-0.42	-0.3	-0.3	-0.03	0.37	-0.14	0.16	-0.23	-0.25	-0.74
YER001W	C26	-2.18	-0.58	0.87	1.71	0.64	0.66	-0.27	-0.43	-0.97	-0.84	0.18	1.46	1.13	1.1	0.31	0.07	-0.86	-0.76
YPR135W	C30	-0.56	-0.76	0.63	1.12	0.51	-0.12	-0.45	-0.79	-0.76	-0.84	0.12	0.57	0.43	-0.29	-0.17	-0.45	-0.42	-0.71
YOL007C	C26	-1.43	-1.25	0.83	0.73	0.77	-0.47	-0.32	-1.18	-1.47	-0.71	-0.32	0.58	0.78	0.33	-0.27	-0.4	-0.84	-1.03
YPL256C	C26	-1.69	-0.97	1.11	1.69	0.45	-0.07	-0.64	-1.6	-1.79	-1.36	0.07	1.29	0.82	0.28	-0.1	-0.6	-0.67	-1.32
YIL140W	C26	-1.43	-1.03	1.37	0.74	0.26	-0.17	-0.84	-1.18	-1.09	-1.03	-0.45	0.7	0.29	-0.36	-0.32	-0.51	-0.6	-1.32
YDR309C	C30	0.53	-0.62	0.33	0.38	0.11	-0.74	-1.09	-1.06	-0.47	-0.3	1.52	0.59	0.64	-0.3	0.53	-0.17	-0.79	-0.42
YMR199W	C26	-1.6	-0.97	1.25	0.83	0.9	0.44	0.03	-0.58	-1.15	-0.81	0.62	1.1	0.95	0.26	0.31	-0.06	-0.45	-0.92
YGR152C	C30	-0.49	-0.58	0.8	0.84	0.57	0.34	-0.01	-0.42	-0.47	-0.38	0.43	0.55	0.42	0.21	0.04	-0.3	-0.17	-0.71
YBL035C	C30	-0.45	-0.64	1.01	1.14	0.45	-0.4	-0.64	0.15	-1.09	0.44	0.04	0.28	0.32	0.03	-0.54	-0.12	-0.6	-0.3
YPR175W	C30	-0.54	-0.69	1.03	0.57	0.49	-0.12	-0.34	-0.62	-0.56	-0.45	0.1	0.52	0.3	-0.22	-0.15	-0.62	-0.2	-0.69
YER111C	C30	-1.25	-0.3	1.32	1.33	0.5	0.14	-0.83	-0.86	-0.79	0.03	0.85	0.74	0.33	-0.23	-0.15	-0.58	-0.38	-0.51

P1	P2	P3	P4	P5	P6	OUT
0.81	0	0.77	0.5	0.72	0.57	0.67

P1 = CORRELAZIONE LINEARE POSITIVA

P2 = CORRELAZIONE LINEARE NEGATIVA

P3 = CORRELAZIONE QUADRATICA

P4 = PERCENTUALE DI GENI CON VALORE FINALE MAGGIORE DEL VALORE INIZIALE

P5 = PERCENTUALE DI GENI CON LO STESSO ANDAMENTO TEMPORALE

P6 = PERCENTUALE DI GENI CON MASSIMA ESCURSIONE NELLO STESSO ISTANTE TEMPORALE

Tab. 9

	alpha 0	alpha 7	alpha 14	alpha 21	alpha 28	alpha 35	alpha 42	alpha 49	alpha 56	alpha 63	alpha 70	alpha 77	alpha 84	alpha 91	alpha 98	alpha 105	alpha 112	alpha 119
YPR120C	0	0.308	0.974	1	0.636	0.456	0.4103	0.297	0.323	0.3333	0.8615	0.815	0.662	0.3846	0.554	0.4	0.39487	0.2513
YJL115W	0.07	0.5	0.564	1	0.548	0.399	0	0.016	0.207	0.2713	0.4202	0.548	0.431	0.1862	0.069	0.1649	0.01596	0.0106
YCR065W	0	0.497	0.884	0.945	0.704	0.648	0.2663	0.377	0.397	0.3116	0.7035	1	0.945	0.804	0.663	0.7538	0.41206	0.4221
YDR097C	0.07	0	0.72	0.979	0.876	0.565	0.114	0.01	0.254	0.0777	0.9585	1	0.777	0.5337	0.415	0.456	0.04663	0.1036
YKL045W	0.19	0.582	0.99	0.981	0.827	0.644	0.3894	0.274	0.192	0.7788	1	0.942	0.803	0.4519	0.534	0.4712	0.37981	0
YNL262W	0.98	0.127	0.759	0.816	1	0.601	0.3354	0.044	0.177	0.1392	0.4684	0.652	0.722	0.5	0.475	0.0949	0.24684	0
YDR074C	0	0.359	0.74	0.961	1	0.74	0.6234	0.173	0.199	0.4459	0.329	0.814	0.81	0.5325	0.377	0.3983	0.22078	0.1472
YER070W	0	0.24	0.658	1	0.747	0.652	0.2601	0.267	0.145	0.3108	0.6115	0.916	0.74	0.5608	0.493	0.5338	0.31419	0.2534
YLR103C	0	0.286	1	0.896	0.727	0.461	0.2208	0.195	0.11	0.1948	0.6753	0.792	0.63	0.3182	0.253	0.3182	0.12338	0.1688
YNL312W	0.06	0	0.726	1	0.897	0.891	0.4743	0.183	0	0.1314	0.4657	0.583	0.754	0.3657	0.486	0.0971	0.32571	0.1429
YJL074C	0.15	0	0.717	1	0.92	0.519	0.4292	0.127	0.142	0.3585	0.5566	0.802	0.797	0.4198	0.373	0.2877	0.29717	0.4057
YJL187C	0.4	0.514	0.751	0.968	0.917	0.719	0.6877	0.668	0.589	0.4743	0.8577	1	0.996	0.8459	0.881	0.8221	0	0.5731
YBR088C	0.03	0.133	0.86	1	0.828	0.488	0.2807	0.716	0	0.2175	0.6526	0.888	0.814	0.5263	0.446	0.2526	0.15439	0.1193
YNL102W	0.02	1	0.312	0.596	0.465	0.184	0.1667	0.174	0.213	0.0177	0.344	0.504	0.496	0.273	0.316	0.0532	0	0.078
YKL113C	0	0.351	0.738	1	0.743	0.565	0.2932	0.173	0.136	0.2147	0.712	0.874	0.848	0.445	0.492	0.4555	0.12042	0.2932
YDL164C	0.12	0.169	0.785	1	0.797	0.441	0.4181	0	0	0.2486	0.5367	0.887	0.814	0.3616	0.333	0.1356	0.15819	0.1356
YGL038C	0	0.43	0.913	0.96	0.336	0.617	0.1141	0.289	0.295	0.7114	0.7919	1	0.785	0.6711	0.51	0.4966	0.27517	0.3624
YPL057C	0.58	0.247	0.934	0.868	0.846	1	0.5659	0.159	0	0.511	0.9286	0.824	0.725	0.5165	0.593	0.2692	0.17582	0.1264
YKL067W	0.13	0.537	0.672	1	0.853	0.944	0.5819	0.35	0.181	0.2486	0.2486	0.401	0.627	0.339	0.508	0.2881	0.27684	0
YER001W	0	0.411	0.784	1	0.725	0.73	0.491	0.45	0.311	0.3445	0.6067	0.936	0.851	0.8432	0.64	0.5784	0.33933	0.365
YPR135W	0.14	0.041	0.75	1	0.689	0.367	0.199	0.026	0.041	0	0.4898	0.719	0.648	0.2806	0.342	0.199	0.21429	0.0663
YOL007C	0.02	0.096	1	0.957	0.974	0.435	0.5	0.126	0	0.3304	0.5	0.891	0.978	0.8087	0.522	0.4652	0.27391	0.1913
YPL256C	0.03	0.236	0.833	1	0.644	0.494	0.3305	0.055	0	0.1236	0.5345	0.885	0.75	0.5948	0.486	0.342	0.32184	0.1351
YIL140W	0	0.143	1	0.775	0.604	0.45	0.2107	0.089	0.121	0.1429	0.35	0.761	0.614	0.3821	0.396	0.3286	0.29643	0.0393
YOR309C	0.62	0.18	0.544	0.563	0.46	0.134	0	0.011	0.238	0.3027	1	0.644	0.663	0.3027	0.621	0.3525	0.11494	0.2567
YMR199W	0	0.221	1	0.853	0.877	0.716	0.5719	0.358	0.158	0.2772	0.7789	0.947	0.895	0.6526	0.67	0.5404	0.40351	0.2386
YGR152C	0.14	0.084	0.974	1	0.826	0.677	0.4516	0.187	0.155	0.2129	0.7355	0.813	0.729	0.5935	0.484	0.2645	0.34839	0
YBL035C	0.29	0.202	0.942	1	0.691	0.309	0.2018	0.556	0	0.6861	0.5067	0.614	0.632	0.5022	0.247	0.435	0.21973	0.3543
YPR175W	0.09	0	1	0.733	0.686	0.331	0.2035	0.041	0.076	0.1395	0.4593	0.703	0.576	0.2733	0.314	0.0407	0.28488	0
YER111C	0	0.368	0.996	1	0.678	0.539	0.1395	0.151	0.178	0.4961	0.814	0.771	0.612	0.3953	0.426	0.2597	0.33721	0.2868

Tab. 10

	|Δ1|	|Δ2|	|Δ3|	|Δ4|	|Δ5|	|Δ6|	|Δ7|	|Δ8|	|Δ9|	|Δ10|	|Δ11|	|Δ12|	|Δ13|	|Δ14|	|Δ15|	|Δ16|	|Δ17|
YPR120C	0.30769	0.66667	0.02564	0.3641	0.17949	0.04615	0.11282	0.02564	0.01026	0.52821	0.04615	0.15385	0.27692	0.16923	0.15385	0.00513	0.14359
YJL115W	0.43085	0.06383	0.43617	0.45213	0.14894	0.39894	0.01596	0.19149	0.06383	0.14894	0.12766	0.11702	0.24468	0.11702	0.09574	0.14894	0.00532
YCR065W	0.49749	0.38693	0.0603	0.24121	0.05528	0.38191	0.11055	0.0201	0.08543	0.39196	0.29648	0.05528	0.1407	0.1407	0.09045	0.34171	0.01005
YDR097C	0.06736	0.72021	0.25907	0.10363	0.31088	0.45078	0.10363	0.24352	0.17617	0.88083	0.04145	0.2228	0.24352	0.11917	0.04145	0.40933	0.05699
YKL045W	0.38942	0.40865	0.00962	0.15385	0.18269	0.25481	0.11538	0.08173	0.58654	0.22115	0.05769	0.13942	0.35096	0.08173	0.0625	0.09135	0.37981
YNL262W	0.85443	0.63291	0.05696	0.18354	0.39873	0.26582	0.29114	0.13291	0.03797	0.32911	0.18354	0.06962	0.22152	0.02532	0.37975	0.1519	0.24684
YOR074C	0.35931	0.38095	0.22078	0.03896	0.25974	0.11688	0.45022	0.02597	0.24675	0.11688	0.48485	0.00433	0.27706	0.15584	0.02165	0.17749	0.07359
YER070W	0.23986	0.61824	0.14189	0.25338	0.09459	0.39189	0.00676	0.12162	0.16554	0.30068	0.30405	0.17568	0.17905	0.06757	0.04054	0.21959	0.06081
YLR103C	0.28571	0.71429	0.1039	0.16883	0.26623	0.24026	0.02597	0.08442	0.08442	0.48052	0.11688	0.16234	0.31169	0.06494	0.06494	0.19481	0.04545
YNL312W	0.05714	0.72571	0.27429	0.10286	0.00571	0.41714	0.29143	0.18286	0.13143	0.35429	0.09714	0.17143	0.38857	0.12	0.38857	0.22857	0.18286
YJL074C	0.15094	0.71698	0.28302	0.08019	0.40094	0.08962	0.30189	0.01415	0.21698	0.19811	0.24528	0.00472	0.37736	0.04717	0.08491	0.00943	0.10849
YJL187C	0.11858	0.23715	0.21739	0.05138	0.19763	0.03162	0.01976	0.07905	0.11462	0.3834	0.14229	0.00395	0.1502	0.03557	0.05929	0.82213	0.57312
YBR088C	0.10175	0.72632	0.14035	0.17193	0.34035	0.20702	0.43509	0.71579	0.21754	0.43509	0.23509	0.07368	0.28772	0.0807	0.19298	0.09825	0.03509
YNL102W	0.97518	0.68794	0.28369	0.13121	0.28014	0.01773	0.00709	0.03901	0.19504	0.32624	0.15957	0.00709	0.2234	0.04255	0.26241	0.05319	0.07801
YKL113C	0.35079	0.38743	0.26178	0.25654	0.17801	0.27225	0.12042	0.03665	0.07853	0.49738	0.1623	0.02618	0.40314	0.04712	0.03665	0.33508	0.17277
YDL164C	0.0452	0.61582	0.21469	0.20339	0.35593	0.0226	0.41808	0	0.24859	0.28814	0.35028	0.07345	0.45198	0.02825	0.19774	0.0226	0.0226
YGL038C	0.42953	0.48322	0.04698	0.62416	0.26188	0.50336	0.1745	0.00671	0.41611	0.08054	0.20805	0.21477	0.11409	0.16107	0.01342	0.22148	0.08725
YPL057C	0.33516	0.66681	0.06593	0.02198	0.15385	0.43407	0.40659	0.15934	0.51099	0.41758	0.1044	0.0989	0.20879	0.07692	0.32418	0.09341	0.04945
YKL067W	0.40678	0.13559	0.32768	0.14689	0.0904	0.36158	0.23164	0.16949	0.0678	0	0.15254	0.22599	0.28814	0.16949	0.22034	0.0113	0.27684
YER001W	0.41131	0.37275	0.21594	0.27506	0.00514	0.23907	0.04113	0.13882	0.03342	0.26221	0.32905	0.08483	0.00771	0.20308	0.0617	0.23907	0.02571
YPR135W	0.10204	0.70918	0.25	0.31122	0.32143	0.16837	0.17347	0.01531	0.04082	0.4898	0.22959	0.07143	0.36735	0.06122	0.14286	0.01531	0.14796
YOL007C	0.07826	0.90435	0.04348	0.01739	0.53913	0.06522	0.37391	0.12609	0.33043	0.16957	0.3913	0.08696	0.16957	0.28696	0.05652	0.1913	0.08261
YPL256C	0.2069	0.5977	0.16667	0.35632	0.14943	0.16379	0.27586	0.0546	0.12356	0.41092	0.35057	0.13506	0.15517	0.1092	0.14368	0.02011	0.18678
YIL140W	0.14286	0.85714	0.225	0.17143	0.15357	0.23929	0.12143	0.03214	0.02143	0.20714	0.41071	0.14643	0.23214	0.01429	0.06786	0.03214	0.25714
YDR309C	0.44061	0.36398	0.01916	0.10345	0.32567	0.1341	0.01149	0.22605	0.06513	0.69732	0.35632	0.01916	0.36015	0.31801	0.2682	0.23755	0.14176
YMR199W	0.22105	0.77895	0.14737	0.02456	0.1614	0.14386	0.21404	0.2	0.1193	0.50175	0.16842	0.05263	0.24211	0.01754	0.12982	0.13684	0.16491
YGR152C	0.05806	0.69032	0.02581	0.17419	0.14839	0.22581	0.26452	0.03226	0.05806	0.52258	0.07742	0.08387	0.13548	0.10968	0.21935	0.08387	0.34839
YBL035C	0.0852	0.73991	0.0583	0.30942	0.38117	0.10762	0.35426	0.55605	0.6861	0.17937	0.10762	0.01794	0.13004	0.25561	0.18834	0.21525	0.13453
YPR175W	0.08721	1	0.26744	0.04651	0.35465	0.12791	0.16279	0.03488	0.06395	0.31977	0.24419	0.12791	0.30233	0.0407	0.27326	0.24419	0.28488
YER111C	0.36822	0.62791	0.00388	0.32171	0.13953	0.39922	0.01163	0.02713	0.31783	0.31783	0.04264	0.15891	0.21705	0.03101	0.16667	0.07752	0.05039